MODELO CONCEPTUAL PARA DETERMINAR EL PRÓXIMO PRESIDENTE DE ESTADO A TRAVÉS DE *DATAMINING* Y *TWITTER*

DANIEL EDUARDO CORZO SOSA

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
ALTERNATIVA INVESTIGACIÓN TECNOLÓGICA
BOGOTÁ
2018

MODELO CONCEPTUAL PARA DETERMINAR EL PRÓXIMO PRESIDENTE DE ESTADO A TRAVÉS DE *DATAMINING* Y *TWITTER*

DANIEL EDUARDO CORZO SOSA

Trabajo de grado para optar al título de
Ingeniero de Sistemas

Director
Ing. John Alexander Velandia Vega M.Sc.

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
ALTERNATIVA INVESTIGACIÓN TECNOLÓGICA
BOGOTÁ
2018

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

_____

_____
Firma del presidente del Jurado

_____
Firma del presidente del Jurado

_____
Firma del presidente del Jurado

Bogotá, 03 de diciembre de 2018

CONTENIDO

TABLA DE FIGURAS

TABLA DE TABLAS

TABLA DE ANEXOS

GLOSARIO

AZURE: Servicio de computo en la nube ofrecido por Microsoft.

DATAMINING: Proceso de analizar un conjunto de datos en busca de patrones que permitan recolectar información útil.

ENDPOINT: Es el identificador de un canal de comunicación que permite a una aplicación ejecutar un proceso o consumir un servicio.

ETL: Siglas de Extract-Transform-Load (extraer-transformar-cargar), referentes a las operaciones de base de datos que permiten manipular datos que están en un formato diferente a la estructura destino para ser cargados.

KDD: Knowledge Discovery in Data, en español se refiere a la metodología para descubrimiento de conocimiento en datos.

MVVM: Patrón de arquitectura que permite la separación de capas de una aplicación entre un modelo, vista y una vista-modelo. Sirve para simplificar una solución que contenga componentes en distintas plataformas.

PDC: Prueba de Concepto, es una herramienta que permite realizar la valoración de un concepto antes de empezar a desarrollarlo por completo.

PLN: Procesamiento de lenguaje natural, hace referencia al área del aprendizaje de máquina que analiza el lenguaje humano, buscando cerrar la brecha existente de comunicación entre hombre y máquina.

POLARIDAD: Numero entre 0 y 1 que indica si un texto es negativo, neutral o positivo.

RBC: Reconocimiento basado en casos, es una metodología usada para resolver nuevos problemas basado en soluciones similares implementadas con anterioridad.

RETWEET: Acto de referenciar en la línea de tiempo propia un tweet de otro usuario.

TWEET: Frase de no más de 280 caracteres que expresa la opinión de un usuario frente a un tema en particular.

TWITTER: Red social que permite interactuar por medio de tweets entre usuarios de la misma red.

RESUMEN

El modelo propuesto en este documento, permite consolidar la polarización de los comentarios generados por usuarios a publicaciones de candidatos a la presidencia de Colombia para el periodo 2018-2022 de la red social twitter, haciendo uso de la API de servicios cognitivos para análisis de texto provista por Azure. Esto indica que tan aceptado o rechazado un candidato es basado en el sentimiento general expresado en comentarios.

Sumado a lo anterior, los resultados del análisis son expuestos por medio de un endpoint Rest, el objetivo de lo anterior es habilitar esa información consolidada a clientes desarrollados en diferentes lenguajes, plataformas y/o sistemas operativos, para que hagan uso de los datos, o bien sea, los extiendan y se genere nuevo conocimiento.

Palabras clave: análisis de sentimientos, azure textmining, datamining, KDD, PLN, polaridad, tweepy, twitter.

INTRODUCCIÓN

En el año 2007, el entonces senador de los Estados Unidos Barack Obama, por casualidad anunció la creación de un comité para valorar sus opciones a la presidencia, ese mismo día, un universitario que simpatizaba con la causa de Obama creó un grupo en Facebook llamado "*One Million Strong for Barack*", al cabo de un mes, ese grupo tenía algo más de 250.000 usuarios asociados. (Jose Antonio Vargas, 2007)

Este comportamiento fue identificado por los asesores, los cuales vieron un nicho que podían explotar en su beneficio, idearon una estrategia a gran escala basada en tener presencia digital en las plataformas importantes de su época, tales como YouTube, MySpace, Flickr, Facebook, Twitter y LinkedIn. Esta estrategia tenía una ventaja muy importante, tener un perfil en esas redes no costaba dinero comparado a los métodos tradicionales de publicidad (TV, Radio, Vallas, Volantes, entre otros) y permitía una masificación del mensaje, incluso más allá del ámbito americano el cual era su objetivo. (Sasha Issenberg, 2012)

Durante la campaña presidencial del 2008, el equipo de Obama logró reunir más de 1.5 millones de usuarios, ahorrar más de 47 millones de dólares[1] y cambió la forma de ver las redes sociales como un lugar donde se expone un producto a una herramienta en potencia para compartir su mensaje, un punto de organización y encuentro entre simpatizantes diferentes entre sí con un único objetivo: Posicionar a Obama como presidente de los Estados Unidos de América.

Para la campaña del 2012, entre los cuatro años de su primer mandato, las redes sociales se masificaron por completo, ya no eran algo que usaban solo cierto grupo de personas, sino era algo que estaba presente en la vida diaria de las personas. El número de seguidores del entonces presidente pasó de 200 mil a varios millones, su repercusión como figura pública era tanto o más inclusiva que un artista, una marca, una celebridad, entre otros.

Es posible decir que el 2012 fue el año de las elecciones del Big Data, ya que el equipo de campaña de Obama, que estaba conformado por matemáticos, programadores y expertos en marketing digital analizaron por cada estado cuáles eran las intenciones de voto de las personas de acuerdo a sus reacciones en las publicaciones realizadas en redes sociales, usando herramientas tecnológicas para análisis de datos tales como *Hp Vertica MPP*, *R*, *Stata*, (Andrew Lampitt, 2013) entre otros.

[1] Miller Claire, How Obama's internet campaign changed politics, [en línea]. The New York Times [citado el 23 noviembre, 2018]. Disponible en <https://bits.blogs.nytimes.com/2008/11/07/how-obamas-internet-campaign-changed-politics/>

De ese modo no solo tenían las intenciones de voto a favor, tenían un numero de los indecisos, lo cual permitiría crear campañas explícitas a ese grupo de personas para convencerlas de dar su voto a favor.

Lo anterior abriría al mundo el siguiente paso en el uso de las redes sociales como una herramienta imprescindible para lograr los objetivos trazados en campaña para simpatizar con los posibles votantes y personas indecisas. (Nickerson & Rogers, 2013)

1. GENERALIDADES

1.1 ANTECEDENTES

El uso de las redes sociales como fuente de influencia en épocas electorales, comenzó a ser explotado durante la campaña de Obama en el 2008. A raíz del éxito de esas campañas, personalidades políticas de todo el mundo han tomado ejemplo e implementado estos mismos modelos. Para un caso específico de uso de twitter como herramienta para obtener simpatizantes, está la campaña de Donald Trump, el cual es el actual presidente de los Estados Unidos. (Alang, 2016)

La captura y posterior análisis de los datos de la campaña de Obama, fueron hechos mediante una herramienta construida por su equipo de ingenieros, llamada Project Houdini (Gallagher, 2012), el cual es una evolución de ORCA, un software que permite habilitar voluntarios en puestos de votación a través del país para reportar cuales votantes se han retractado o están indecisos, eso con el fin de realizar campañas específicas para atraer esos votos a favor del candidato. (Jacobs, 2012).

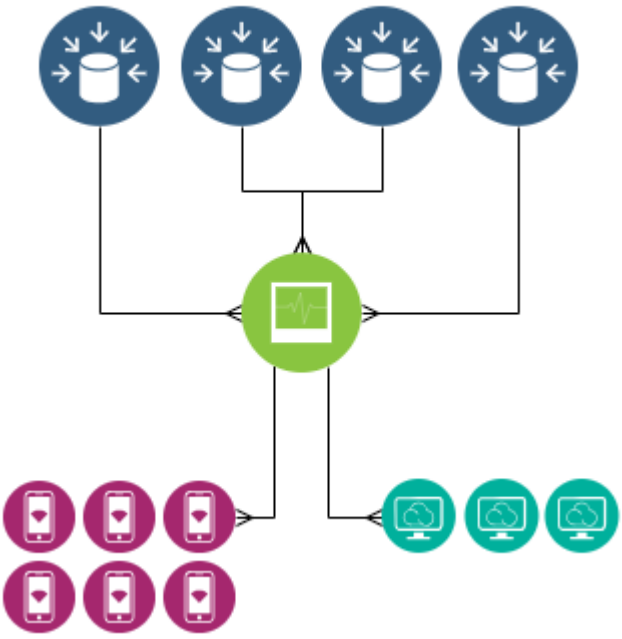
ORCA realmente fue concebido como la herramienta clave para atraer votantes en la campaña del rival de Obama: Mitt Romney, fue hecha por voluntarios simpatizantes con la causa de Romney, sin embargo, no fue hecha teniendo en cuenta las mejores prácticas para proyectos de IT, eso generó que los análisis presentados al comité que acompañaba las decisiones del entonces candidato, fueran errados, los cuales los llevaron a tomar una serie de acciones en pro de captar votos, sin saber que estaban haciendo todo lo contrario. Fue blanco de sabotaje, ataques de denegación de servicio, fallas constantes en el sistema. (Sean Gallagher, 2012).

Las malas decisiones fueron intervenir en lugares donde Obama tenía la mayoría de votantes a su favor, hacer campañas de atracción de votos en zonas donde no había mayor afluencia o simplemente no tener en cuenta algunos otros lugares del país donde existían grandes poblaciones que estaban neutrales en su decisión de voto. Todo lo anterior gracias a los graves fallos que ORCA tenía, pero nadie había tenido en cuenta.

Su principal defecto fue su arquitectura, ya que no fue pensada teniendo en cuenta la gran afluencia de votantes y usuarios. La aplicación web que gestionaba todas las métricas estaba soportada en un único servidor de aplicaciones conectado a 11 servidores de bases de datos.

El sistema no le fue ejecutado un plan de pruebas lo suficientemente exhaustivo para determinar su comportamiento una vez estuviera disponible al público en general, tampoco fue impartida la suficiente capacitación a los usuarios para manejar la herramienta en sus versiones Web y consola. Esto hacia que la aplicación no estuviera disponible ya que todo el tráfico confluía a un solo punto y se generaban cuellos de botella en las peticiones y respuestas desde y hacia su único servidor de aplicaciones.

Figura 1 Arquitectura de Orca



Fuente (AUTOR, 2018)

El equipo de ingeniería de Obama, teniendo en cuenta los fallos presentados por ORCA y sus evidentes malos resultados a la campaña de Romney, plantearon un sistema que hiciera lo mismo: *Project Houdini*, pero sin los errores de diseño, incluyendo modelos estadísticos, probabilísticos y *big data* desde el inicio, con una metodología de trabajo que tuviera en cuenta un amplio rango de pruebas funcionales, de carga, estáticas y dinámicas, que permitirían anticipar posibles problemas antes de ser lanzada al público, lo cual, aseguraría márgenes de resultados más precisos para la toma de decisiones.

Project Houdini le dio una ventaja enorme a Obama respecto a su contrincante Romney en 2008, gracias al análisis de los datos capturados por los voluntarios que estaban de acuerdo con las propuestas de Obama. Se dejaron de usar canales telefónicos para la transmisión de datos (que son lentos), en vez de eso, se usó internet y su masificación a nivel global, lo cual sumado a una arquitectura en la nube y una serie de buenas prácticas de desarrollo al software implementado permitió focalizar esfuerzos en la atracción de votos, a lo cual se le consideraría un movimiento revolucionario dentro de la industria del software en pro de la política, ya que Obama sería declarado como el primer presidente que ganó usando big data.

De ese modo, las técnicas estadísticas, de minería de datos y de big data, no fueron vistas solo para realizar análisis de datos o bajo el contexto académico, sino como herramientas poderosas para lograr objetivos sobre áreas prácticas que no tienen relación directa con la ingeniería de la información.

En Colombia, es visible el uso que los candidatos a la presidencia le están dando a las redes sociales, ya que han visto la efectividad de las campañas gracias a los resultados de otros candidatos de otros países. la masificación del acceso a internet y con ello, el activismo de los simpatizantes y su influencia sobre subgrupos de personas que se encargan de compartir contenidos, expandiendo el mensaje que transmiten. (Semana, 2018)

Actualmente, la empresa colombiana ResolveStudio tiene disponible una herramienta privada que permite mostrar el resultado del análisis que realizan sobre 12 millones de datos en redes sociales, búsquedas y reacciones, el cual indica el índice de intención de voto hacia los candidatos a la presidencia de Colombia. (ResolveStudio, 2018).

1.2 PLANTEAMIENTO DEL PROBLEMA

1.2.1 Descripción del problema.

Durante el periodo en el que los candidatos presidenciales ejecutan sus respectivas campañas, se basan en estrategias elaboradas por sus equipos para consolidar la aceptación de las masas a su favor, lo cual en algunas ocasiones es bien visto por unos pero rechazada por otros.

Hoy en día, un candidato no solo recurre a su oratoria para influir sobre la población, ni a cuantos contactos de personalidades clave dentro del organigrama político del país tenga, las redes sociales están jugando un papel decisivo al momento de realizar acciones, ya que, al estar presentes en muchos aspectos de la vida diaria de una persona, permiten influir en cuales podrían ser sus futuras decisiones.

Las redes sociales, más allá de otorgar libre acceso a una persona a un sinfín de artículos de su interés, se usan como fuente de influencia hacia un usuario para persuadirlo que vote por cierto candidato, bien sea valiéndose de información veraz o de información falsa, generando una serie de reacciones expresadas en comentarios, distribuciones recurrentes de las publicaciones entre la red de quien lee, emoticones, entre otras acciones, sin embargo, ¿cómo un modelo conceptual podría medir la información existente en una red social, específicamente en twitter , para inferir cual candidato tendría más posibilidades de quedarse con la presidencia de Colombia?.

1.2.2 Formulación del problema

Las redes sociales brindan varios métodos que permiten indicar la reacción de un usuario ante un recurso publicado bien sea por un candidato o por un tercero, que simpatice o rechace las propuestas dadas, sin embargo, no está al alcance inmediato, una herramienta tecnológica que permita medir esa reacción indicada en comentarios y *likes*, por lo tanto, ¿Cómo podría medirse que tan aceptado es un candidato a comparación a sus contendientes, basado en las reacciones de la gente hacia las publicaciones en twitter relativas a sus ideas de campaña?

Se plantea el uso de técnicas de *datamining* que permitan manipular la información extraída de *twitter* en busca de patrones que expliquen el comportamiento de los usuarios ante una publicación.

1.3 OBJETIVOS

1.3.1 Objetivo General.

Implementar un modelo que infiera el próximo presidente de Colombia, basado en las reacciones de usuarios a publicaciones en twitter.

1.3.2 Objetivos específicos.

- Seleccionar una técnica de *Datamining* que permita modelar los datos recopilados en *Twitter.*

- Diseñar la estructura del modelo respecto al resultado de la fase de análisis.

- Desarrollar una prueba de concepto web, que tenga en cuenta los módulos definidos en la estructura del modelo.

- Ejecutar un conjunto de pruebas funcionales que permitan evaluar la calidad de los resultados calculados por la solución.

1.4 JUSTIFICACIÓN

Las herramientas que son usadas por los equipos de estrategias digitales en campañas presidenciales, no son abiertas al público, son programas realizados por empresas privadas, las cuales tienen un interés con ánimo de lucro. La propuesta del modelo que presenta este documento, pretende ser abierta a la comunidad académica la cual pueda realizar aportes para mejorar los elementos planteados, a su vez que sea un punto de partida para aquellas personas que quieran incursionar en el mundo del *Datamining* desde un punto de vista práctico.

El análisis de publicaciones respecto a la opinión que tienen las personas hacia un candidato presidencial, es el motivador del planteamiento del modelo conceptual en este documento. No solo permitirá a un ingeniero realizar la abstracción del problema a un sistema de información, también es un punto de entrada al área de la minería de datos, que hoy en día, se presenta como una de las herramientas que complementan el componente técnico y teórico de un profesional de ingeniería.

1.5 ALCANCES Y LIMITACIONES

1.5.1 Alcances.

- El despliegue de la solución permitirá graficar los resultados de la polarización de los tweets usando el modelo propuesto.

- El tiempo de desarrollo será menor o igual a 16 semanas.

- Solo abarcará las reacciones sobre las cuentas de usuario de los candidatos presidenciales colombianos del periodo 2018-2022.

1.5.2 Limitaciones.

- La implementación del modelo se llevará a cabo en el segundo semestre académico del 2018.

- El tiempo de desarrollo estará limitado al periodo académico en curso.

- Se usará la API de Azure para analítica de texto en su versión gratuita.

- Se usará la API de búsqueda de tweets en su versión estándar, la cual tiene como restricción que solo se pueden consultar los últimos 3200 tweets de cada candidato.

1.6 MARCO DE REFERENCIA

1.6.1 Marco teórico.

El análisis de datos es un término nuevo para muchas personas. Hace algunos años, la información almacenada no tenía el sentido que tiene hoy en día, ya que, a partir del análisis ejecutado sobre ella, es posible tomar decisiones que permitan tomar una ventaja competitiva según el mercado al que pertenezcan (Sanon, 2017).

Conociendo lo anterior, se presentan algunos conceptos clave para entender como la información podría ser transformada de modo que permita dar una visión más amplia de su objetivo final.

1.6.1.1 Datamining.

Es el proceso de encontrar anomalías, patrones y correlaciones en grandes volúmenes de datos para predecir resultados. Puede ser usado para incrementar ingresos, disminuir costos, mejorar las relaciones con el cliente, reducir riesgos, entre otros elementos más. (SAS, 2018)

Es también conocido como descubrimiento de conocimiento en bases de datos, en el área de las ciencias de la computación hace referencia al proceso de descubrir interesantes y útiles patrones y relaciones en grandes volúmenes de datos. Este campo combina herramientas como la estadística e inteligencia artificial (tales como redes neurales y aprendizaje de maquina) con administración de bases de datos para analizar grandes volúmenes conocidos como datasets. (Clifton, 2018)

La minería de datos es ampliamente usada en negocios (aseguradoras, banca), investigación científica (astronomía, medicina) y seguridad gubernamental (detección de criminales y terroristas)

- Origen y primeras aplicaciones

Como la capacidad de almacenamiento de los computares ha aumentado desde los años 80, muchas compañías comenzaron a guardar mucha más información transaccional. Las colecciones resultantes, comúnmente llamadas *datawarehouses* (Amazon, 2018), eran muy grandes para ser analizadas con métodos tradicionales estadísticos.

Algunas conferencias y talleres se llevaron a cabo para considerar como los recientes avances en el campo de la inteligencia artificial, tales como los descubrimientos de sistemas expertos (Zwass, 2018), algoritmos genéticos, aprendizaje de máquina (Faggella, 2017) y redes neurales, podrían ser adaptados para descubrimiento de conocimiento.

Este proceso comenzó en 1995 en Montreal como la primera conferencia internacional sobre descubrimiento de conocimiento y minería de datos. (aaai.org, 1995)

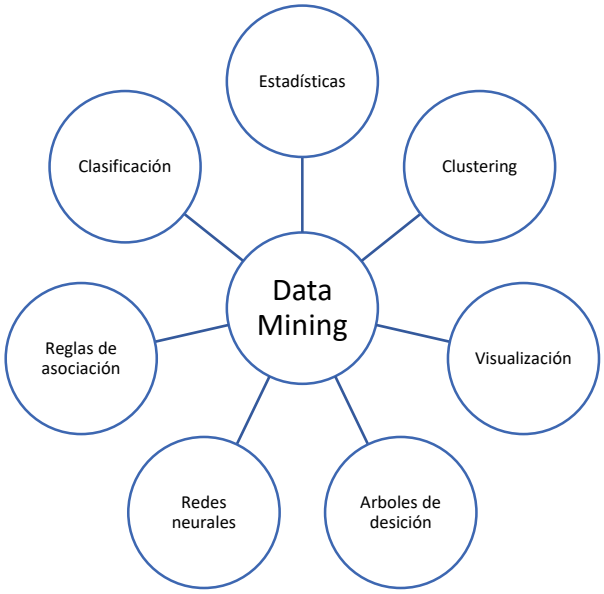
Una de las primeras aplicaciones del datamining, tal vez de segundo únicamente por la investigación de marketing, fue la detección de fraudes de tarjetas de crédito (Akhilomen, 2013). Estudiando el comportamiento de compras de un consumidor, un patrón usual aparece, si aparecen transacciones realizadas fuera de ese patrón, se marcan como en investigación para ser aprobadas o rechazadas en procesos de verificación posteriores, sin embargo, la amplia variedad de comportamientos normales puede hacer de esta tarea algo desafiante, no siempre las distinciones entre una transacción normal y una riesgosa aplican a todas las personas. (Bhattacharyya, 2011)

Un ejemplo de lo mencionado anteriormente podría darse como los viajeros de negocios frecuentes, ellos podrían realizar compras en distintas partes del mundo por montos diversos, lo cual podría catalogarse como un fraude, sin embargo, este caso no clasifica dentro del grupo de transacciones marcadas como rechazo por una plataforma de detección de fraudes. (Clifton, 2018)

- Técnicas de *Datamining*

Hay varios tipos de minería de datos, típicamente dividido por el tipo de información conocida y el tipo de conocimiento buscado del modelo usado.

Figura 2 Técnicas de Datamining



Fuente (AUTOR, 2018)

- Estadísticas

Las técnicas de minería de datos estadísticas son una rama de las matemáticas que describe las colecciones y descripciones de datos. Las técnicas estadísticas no son consideradas como parte de una técnica de minería de datos por varios analistas, pero ayuda a descubrir patrones y construir modelos predictivos (Shethna, 2017). Los analistas de datos deberían poseer algunos conocimientos con respecto a las diferentes técnicas estadísticas, ya que permitirá concluir respuestas a preguntas como las siguientes:

- ¿Cuáles son los patrones en la base de datos?
- ¿Cuál es la probabilidad de ocurrencia de un evento?
- ¿Cuáles patrones son más útiles para el negocio?
- ¿Cuál es el resumen de alto nivel que puede darle una vista detallada de lo que hay en la base de datos?

La estadística ayuda a resumir la información y contarla, además de proveer información respecto a los datos con facilidad. A través de los análisis estadísticos, las personas pueden tomar decisiones inteligentes. (Shethna, 2017). Hay distintas formas de técnicas estadísticas, tales como:

- Histogramas
- Mediana
- Moda
- Varianza
- Máximos
- Mínimos
- Regresión lineal

- Clustering

Es una de las técnicas más antiguas usadas en minería de datos. Este es el proceso de identificar datos que son similares a otros, esto ayudaría a entender las diferencias y similitudes entre los datos. (CrayonData, 2015) Esto es conocido algunas veces como segmentación y ayuda al usuario a entender que está ocurriendo en la base de datos.

Existen principalmente dos tipos diferentes de técnicas de clustering:

- Clustering jerárquico

Esta técnica construye un árbol que representa las similitudes entre los distintos elementos. La exploración de todos los posibles arboles es computacionalmente intratable, por lo que se suelen usar algoritmos aproximados guiados por heurísticas (Berzal, 2017). Dentro de esta clasificación se pueden generar 2 subgrupos

➢ Clustering jerárquico aglomerativo

Se comienza con tantos clústeres como individuos y consiste en ir formando (aglomerando) grupos según su similitud.

➢ Clustering jerárquico de división

Se comienza con un único clúster y consiste en ir dividiendo clústeres según la diferencia entre sus componentes.

- Clustering de partición

Esta técnica realiza una distribución de los elementos entre un numero prefijado de clústeres o grupos. Recibe como dato de entrada el número de clústeres a formar además de los elementos a clasificar y la matriz de similitudes (Romero-Campero, 2013). Explorar todas las posibles particiones es computacionalmente intratable. Por lo tanto, suelen seguirse algoritmos aproximados guiados por determinadas heurísticas. En lugar de construir un árbol el objetivo en PAM consiste en agrupar los elementos entorno a elementos centrales llamados centroides a cada clúster.

- Visualización

Es la técnica más usada en cuanto a descubrimiento de patrones de datos. Permite convertir datos pobres a buenos datos permitiendo la aplicación de diferentes tipos de minería de datos para ser usados en descubrimientos de patrones ocultos (Hernández-Orallo, 2015)

- Arboles de decisión

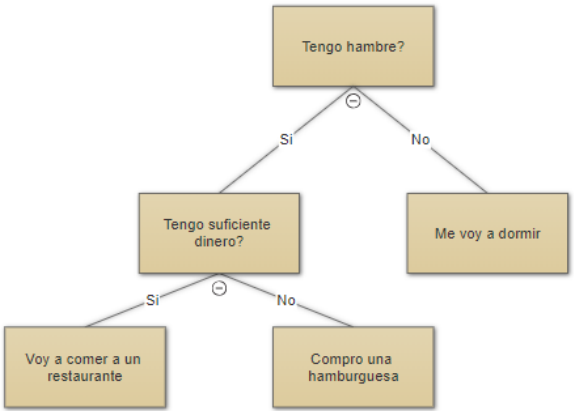
Son modelos predictivos que al ser graficados parecen un árbol. En esta técnica, cada rama del árbol es vista como una pregunta de clasificación y las hojas son consideradas como particiones del dataset relacionado a una clasificación particular. Es ampliamente usado en análisis exploratorio, pre procesamiento de datos y trabajo de predicción. (Gupta, 2017)

El primer paso en esta técnica es hacer crecer el árbol. La base de crecimiento depende en la búsqueda de la mejor pregunta posible a ser preguntada en cada rama del árbol, por lo tanto, este deja de crecer bajo una de las siguientes circunstancias:

- Si un segmento tiene un solo registro
- Todos los registros contienen características idénticas
- El crecimiento no es suficiente para crear una nueva división

Los arboles de decisión proveen resultados que pueden ser fácilmente entendidos por el usuario.

Figura 3 Árbol de decisión



Fuente (AUTOR, 2018)

- Redes neurales

Son una abstracción del modo en el que las células cerebrales (neuronas) procesan información. Es un modelo bioinspirado útil en aplicaciones de clasificación de negocio gracias a su habilidad de aprender de la información que lo compone, su naturaleza no paramétrica y su habilidad para generalizar. (Gaur, 2012)

Hay dos partes importantes dentro de esta técnica:

- Nodo: Representa a la neurona del cerebro humano
- Dendrita: Representa las conexiones entre neuronas del cerebro humano

Una red neural es una colección de neuronas intercomunicadas entre sí, las cuales pueden formar una o múltiples capas. El modo en que se conectan es conocido como arquitectura de la red. Existen diversos modelos y cada uno de ellos tiene sus ventajas y desventajas, cada modelo tiene su propia arquitectura y a su vez cada arquitectura su propio procedimiento de aprendizaje.

Una de las aplicaciones de las redes neurales es en la detección de fraudes en transacciones electrónicas.

- Reglas de asociación

Esta técnica ayuda a encontrar la asociación entre dos o más ítems. Ayuda a conocer las relaciones entre las diferentes variables en bases de datos, descubre patrones ocultos en sets de datos lo cual es usado para identificar variables y la frecuencia de ocurrencia de diferentes variables que aparecen con las más altas frecuencias

- Clasificación

Es una de las técnicas de minería de datos más usadas, las cuales contienen un set de datos preclasificados que sirven para crear el modelo que puede clasificar un volumen amplio de datos. Esta técnica ayuda en la derivación de información importante de la data y de su metadata (Knight, 2017). También está relacionada con el análisis de clustering y usa arboles de decisión o un sistema de redes neurales.

En esta técnica hay dos procesos principales involucrados

- Aprendizaje: En este proceso, la data es analizada por un algoritmo de clasificación.
- Clasificación: Este proceso mide la data respecto a la precisión de las reglas de clasificación.

Algunos de los modelos de clasificación más usados son

- Clasificación por inducción de árboles de decisión
- Clasificación bayesiana
- Redes neurales
- Máquinas de soporte vectorial
- Clasificación basada en asociaciones

Un ejemplo de un clasificador es un proveedor de correo electrónico.

1.6.1.2 Modelos conceptuales.

Explica los conceptos más significativos en un dominio del problema. Muestra los conceptos, asociaciones entre conceptos y atributos de los mismos a partir de palabras o imágenes que lo representan, las proposiciones que expone con claridad sus características o el conjunto de ejemplo al cual se aplica (Ardila, 2018).

1.6.1.3 Pruebas funcionales

Es un tipo de pruebas donde el sistema es probado contra los requerimientos funcionales y/o especificaciones. Las funciones son probadas alimentando las entradas de datos y examinando su salida, lo cual garantiza que el requerimiento está siendo correctamente satisfecho por la aplicación. Este proceso no tiene en cuenta cómo el procesamiento ocurre, sino los resultados entregados por ese procesamiento.

Los pasos involucrados durante las pruebas funcionales son:

- Identificar funciones que se supone que la aplicación debería hacer
- Crear entradas de datos basados en las especificaciones propuestas
- Determinar las salidas de datos basados en las especificaciones propuestas
- Ejecutar casos de prueba
- Comparar los resultados obtenidos contra los esperados

Las ventajas que se encuentran son las siguientes:

- Simula el estado actual de uso del sistema
- No hace ninguna suposición en la estructura del sistema

Las desventajas de usar pruebas funcionales son:

- Las pruebas realizadas pueden tener errores de lógica de quien haya diseñado los casos.
- Alta posibilidad de realizar pruebas redundantes.

Las pruebas funcionales son más eficientes cuando las condiciones de prueba son creadas directamente por los requerimientos de usuario o negocio.

1.6.2 Marco conceptual.

1.6.2.1 SQL

Es un acrónimo para *Structured Query Language* (Lenguaje de consultas estructurado), el cual es usado principalmente para retornar y manipular datos almacenados en una base de datos relacional. Permite manipular grandes conjuntos de datos de manera rápida y eficiente. (Earls, 2010)

1.6.2.2 Datamining

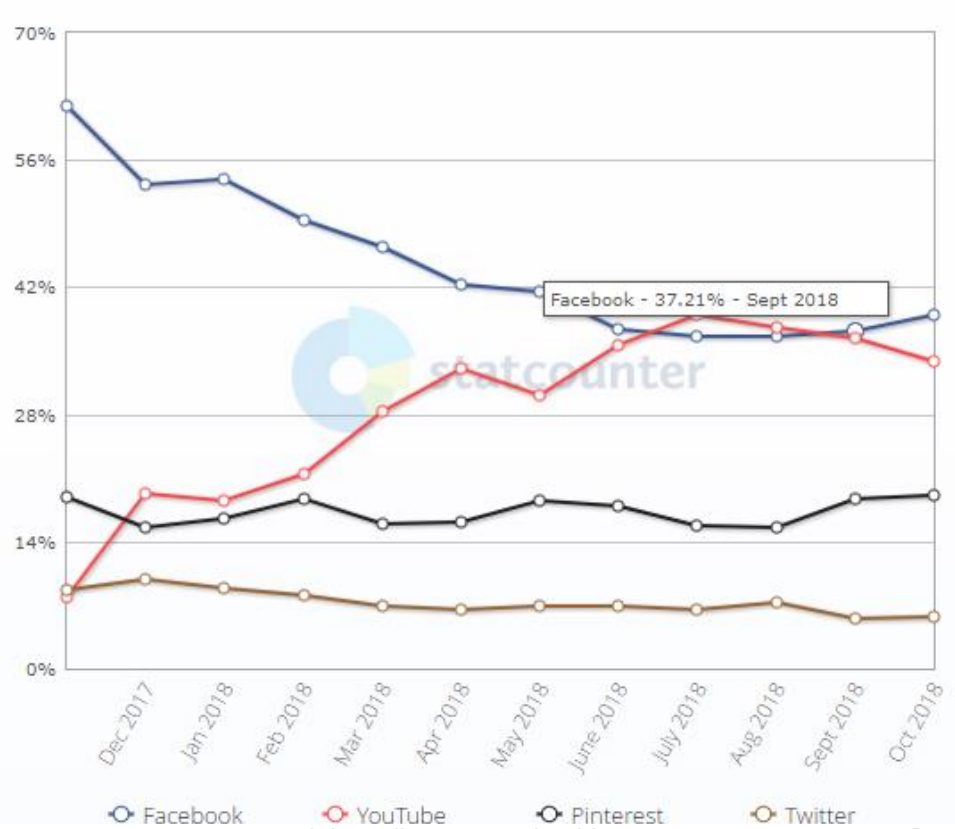
Es la práctica de realizar búsquedas en grandes volúmenes de datos para descubrir patrones y tendencias que van más allá de un simple análisis de datos. Usa métodos matemáticos para segmentar los datos y evaluar la probabilidad de futuros eventos. (Oracle, 2018)

1.6.2.3 Twitter

Red social de noticias donde la gente se comunica usando mensajes cortos llamados tweets (Gil, 2018)

Según las estadísticas de uso de redes sociales provistas por statcounter en el periodo de noviembre de 2017 a octubre de 2018 para Colombia, twitter alcanza un porcentaje de uso del 5.63%[2], lo cual lo ubica en el cuarto puesto de las redes sociales más usadas. Las tres más activas en Colombia son Facebook (38.96%), Youtube(33.75%) y Pinterest(18.98%). El tráfico en twitter corresponde a 16.8 millones de peticiones en el 2018.

Figura 4 Uso de redes sociales en Colombia



Fuente: statcounter, Social media stats Colombia Nov 2017 – Oct 2018. [en línea] Globalstats. [citado el 23 de noviembre, 2018]. Disponible en <http://gs.statcounter.com/social-media-stats/all/colombia/#monthly-201711-201810>

[2] Statscounter, social media stats Colombia. GlobalStats [en línea] Statscounter GlobalStats [citado el 23 noviembre, 2018]. Disponible en < http://gs.statcounter.com/social-media-stats/all/colombia/#monthly-201711-201810>

1.6.2.4 Modelo Conceptual

Permite realizar abstracciones de problemas del mundo real, de modo que se pueda crear un punto de partida hacia el análisis de una problemática y posterior desarrollo de ideas que permitan un acercamiento a una solución. (Aldrich & Garrod, 2013)

1.6.2.5 Metadatos

Es la información con respecto a la data recolectada. Incluye información respecto a criterios técnicos y procesos de negocio, reglas de datos y relaciones. (Knight, 2017)

1.6.2.6 KDD (Knowledge Discovery in data)

Es una metodología usada para identificar patrones válidos, nuevos, potencialmente útiles y comprensibles en un conjunto de datos. No es un proceso automático, es iterativo, permite validar las relaciones entre volúmenes amplios de datos, extrae información que puede usarse para llegar a conclusiones basados en relaciones o modelos dentro de los datos.

Los pasos de la metodología son los siguientes:

- Selección de datos

Es donde se determinan las fuentes de datos y el tipo de información a extraer, buscando los atributos apropiados para representar el objetivo al cual se desea llegar.

- Preprocesamiento

Esa fase remueve los elementos innecesarios de la fuente de datos: ruido y elementos inconsistentes y deja solamente los que son apropiados dentro del análisis que se está realizando.

- Transformación

Permite realizar el tratamiento de los datos y generación de una estructura apropiada para el procesamiento en la siguiente fase. De ser necesario se usan algunas operaciones de agrupamiento o normalización para facilitar el acceso y consulta desde fases posteriores.

- Data Mining

Esta fase es donde los métodos de minería de datos son aplicados con el objetivo de descubrir y extraer patrones desconocidos y potencialmente útiles en el estudio que se anda realizando.

- Interpretación y Evaluación

Una vez se identifican los patrones obtenidos, se aplican medidas y evaluaciones de los resultados obtenidos como conclusión al objeto de estudio que se ejecutó durante todo el proceso de descubrimiento.

1.6.2.7 RBC (Razonamiento basado en casos)

Es un proceso repetitivo para solucionar los problemas actuales mediante los problemas similares ocurridos anteriormente. Básicamente revisa si han existido casos similares y utiliza las soluciones aportadas para intentar solucionar lo mejor posible el problema actual. Esta metodología intenta simular el comportamiento humano utilizando los casos antiguos que ha habido para dar la mejor solución al problema actual[3].

Se pueden mencionar las siguientes ventajas de usar la metodología RBC:

- Reduce las tareas correspondientes a la adquisición de conocimiento.
- Evita repetir errores cometido en el pasado.
- Permite razonamiento en dominios que no han sido completamente entendidos, definidos o modelados.
- Razonamiento en dominios con un conocimiento bajo.
- Puede ser usado en un amplio rango de dominios y formas

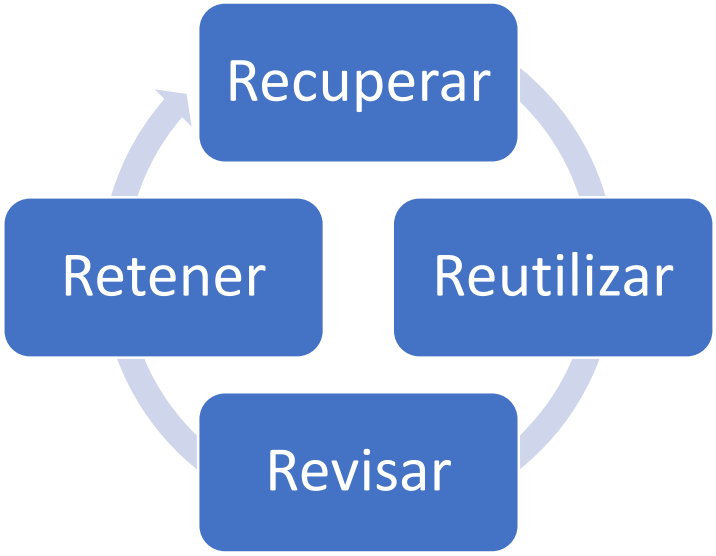
Como desventaja, se puede mencionar lo siguiente:

- El tamaño de la base de datos de casos aumenta siempre, lo cual hace que en algunos casos donde la solución de un problema sea motivo de generación de múltiples soluciones, tanta información tienda a entorpecer a quien usa la metodología.

[3] UDT-IA. Razonamiento basado en casos [en línea]. Unidad de desarrollo tecnológico en Inteligencia Artificial [citado el 23 noviembre, 2018]. Disponible en <http://www.iiia.csic.es/udt/es/artificialintelligence/razonamiento-basado-en-casos>

Dentro del RBC, hay cuatro pilares que se tienen que cumplir desde que se inicia el problema hasta que se obtiene una solución al mismo:

Figura 5 Ciclo de vida RBC



Fuente (AUTOR, 2018)

- Recuperar

En esta fase, el problema actual es cotejado con otros problemas almacenados en una base de casos. Cotejar se refiere a comparar dos casos y determinar su grado de similitud basado en el conocimiento del dominio.

- Reutilizar

Este paso puede darse mediante integración de la solución en los casos que fueron recuperados previamente. Puede darse usando sustitución, transformación y adaptación generativa.

- Revisar

Hace énfasis en evaluar la solución dada por el reúso de uno o varios casos previos. Si es necesario ajustes se realizan en esta fase y se le llama reparación.

- Retener

Una vez la solución propuesta resuelve el problema indicado, se procede al almacenamiento de la misma en una base de casos para su uso futuro, la aprobación depende de lo útil que pueda ser la solución, ya que RBC al ser incremental, necesita depurar su base de casos para no tomar en cuenta aquellas soluciones que ya están incluidas en otras soluciones.

1.6.2.8 Python

Es un lenguaje de programación que ofrece una sintaxis sencilla para realizar programas. Es interpretado, por lo cual permite ahorrar mucho tiempo durante el desarrollo ya que no es necesario compilarlo o empaquetarlo para que funcione. El intérprete puede usarse interactivamente, lo cual facilita escribir programas descartables, realizar pruebas de concepto sencillas o modelar un sistema de información con poco código, siendo compacto y legible.[4]

1.7 METODOLOGÍA

1.7.1 XP (extreme programming).

Es una metodología *agile* para desarrollo de software, que apunta a producir software de alta calidad mayormente en equipos de desarrollo pequeños y cuya tecnología usada permite realizar pruebas unitarias automatizadas y funcionales. (What is extreme programming, 2018)

Esta metodología está basada en mejoras continuas (iteraciones), lo cual permite realizar ajustes al proyecto, resultado de las revisiones sugeridas por el director de proyecto asignado.

[4] Van Rossum Guido, Python tutorial. [en línea] Python Software Foundation.[citado el 23 noviembre, 2018]. Disponible en <http://docs.python.org.ar/tutorial/pdfs/TutorialPython2.pdf>

Los pasos iterativos a realizar durante la ejecución del proyecto se describen en la figura 4

Figura 6 Ciclo de vida de una solución usando XP

Fuente (AUTOR, 2018)

1.7.2 Planeación.

Es la primera fase donde se consolida el equipo de desarrollo para crear historias de usuario o requerimientos, las cuales se convierten en iteraciones que cubren una pequeña parte de la funcionalidad de características requeridas. Una combinación de iteraciones provee al cliente la funcionalidad completa.

1.7.3 Diseño.

Una vez las historias hayan sido convertidas a iteraciones, se aplican los siguientes principios:

- Mantener la simplicidad expresando una única cosa y no adicionando funcionalidades anticipadas.
- Usar nomenclaturas que permitan estandarizar el uso de los recursos a todos los integrantes del equipo de desarrollo.
- Usar un tablero de gestión de historia (Kanban) que permita al equipo compartir sus mejores ideas dentro del diseño.
- Crear soluciones simples o pruebas de concepto que exploren soluciones potenciales para un problema específico.

1.7.4 Codificación

Esta constituye una de las más importantes fases en el ciclo de vida de XP. Dentro de esta metodología se da prioridad a la escritura de código sobre todas las otras tareas tales como documentación en pro que el cliente reciba algo sustancial en valor al final del día.

1.7.5 Pruebas

Esta fase se incluye dentro de la codificación en vez de realizarla al final de todo el desarrollo. Las pruebas unitarias hechas desde un principio garantizan un menor número de bugs y hace que los despliegues gasten menor tiempo en ser ejecutados, garantizando una alta calidad desde la primera entrega.

1.7.6 Retroalimentación

La base de XP es un mecanismo continuo de retroalimentación del cliente durante las iteraciones. Cada retroalimentación del cliente hacia las historias revisadas se convierte en la base de una nueva planeación o en el fin de la misma, lo cual comenzaría de nuevo el ciclo desde planeación hasta una próxima retroalimentación.

2. SELECCIONAR UNA TÉCNICA DE DATAMINING QUE PERMITA MODELAR LOS DATOS RECOPILADOS EN TWITTER

Una de las preguntas que naturalmente vienen al momento de realizar una solución que involucre datamining, es la selección de la técnica que mejor se ajuste al requerimiento propuesto, sin embargo, la respuesta varía básicamente por dos factores (Gibert, Sanchez-Marre, & Codina, 2014):

- El objetivo principal del problema a ser resuelto.
- La estructura de los datos en el dataset.

El contexto del objetivo principal propuesto en este documento, indica que la aceptación o rechazo de un candidato se calculará basado en los comentarios de las personas a las publicaciones realizadas por el candidato en twitter, mientras que la estructura del dataset provisto por la API de twitter, es concisa, presenta siempre los mismos campos y el mismo formato.

Para determinar cuál técnica es la que más se ajusta al objetivo y estructura anteriormente mencionada, es necesario aplicar una metodología que permita tomar esa decisión.

2.1 RBC: Razonamiento basado en casos

Es una metodología cíclica para resolución de problemas que intenta resolver los nuevos, reusando soluciones similares aplicadas en el pasado. Está inspirado en el razonamiento humano y el uso de la memoria (Ortiz, Bañuelos, & Rodas-Osollo, 2016). El detalle de cada fase implementada se encuentra en el marco conceptual, apartado RBC (Ver RBC (Razonamiento basado en casos))

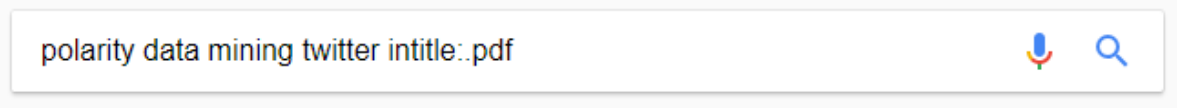
2.2 Ejecución de la metodología en la selección de la técnica

2.2.1 Recuperación

2.2.1.1 Búsqueda usando operadores de google

Google en su buscador, ofrece algunos operadores que permiten generar búsquedas específicas. Para este caso, se usa el operador ***intitle***, el cual filtra los resultados de búsqueda a solo aquellos que son documentos PDF, sumado a palabras clave que identifiquen el contexto del problema planteado en este documento. La figura 6 indica los criterios usados para realiza la búsqueda de artículos.

Figura 7 Recuperación de papers usando google

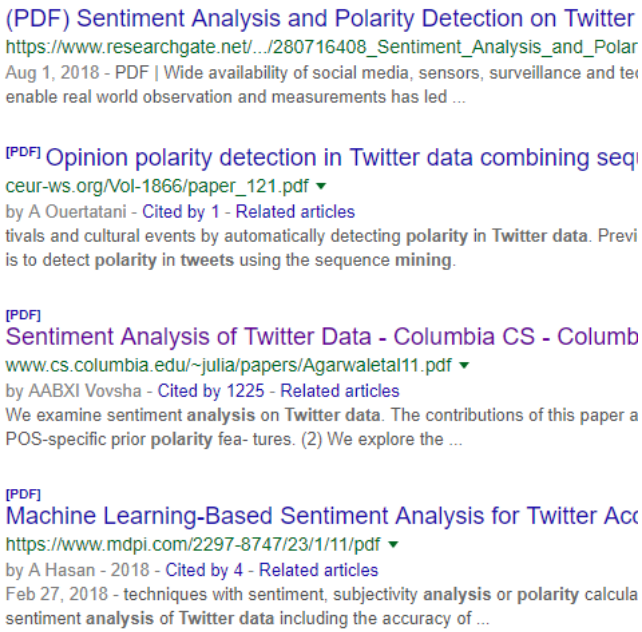


Fuente (AUTOR, 2018)

Se procede a consultar los resultados de búsqueda, en busca de artículos relevantes al uso de twitter y polarización de los mismos, los cuales pueden estar en bases de datos indexadas o expuestos por portales educativos. En total fueron generados 72.400 resultados.

Siguiendo los criterios de búsqueda suministrados y aprovechando la relevancia dada por el buscador de Google, se descargan aquellos que tienen citaciones y vínculos relacionados al portal *Scholar* de Google, al igual los que están en el portal *ResearchGate* y se descartan aquellos vínculos publicitados, que no tienen relevancia en el problema de selección de técnicas de datamining.

Figura 8 Resultados de búsqueda de papers con google

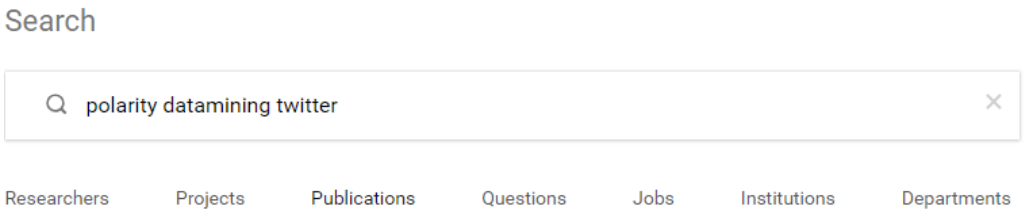


Fuente (AUTOR, 2018)

2.2.1.2 Búsqueda sobre bases de datos indexadas

Las mismas palabras clave usadas para generar búsquedas en google sirven para buscar resultados sobre bases de datos, sin embargo, en este punto no es necesario especificar el tipo de documento a retornar ya que los artículos disponibles se encuentran en formato PDF. Usando *ResearchGate*, se procede a buscar las mismas palabras clave usadas en google, filtrando solo por publicaciones.

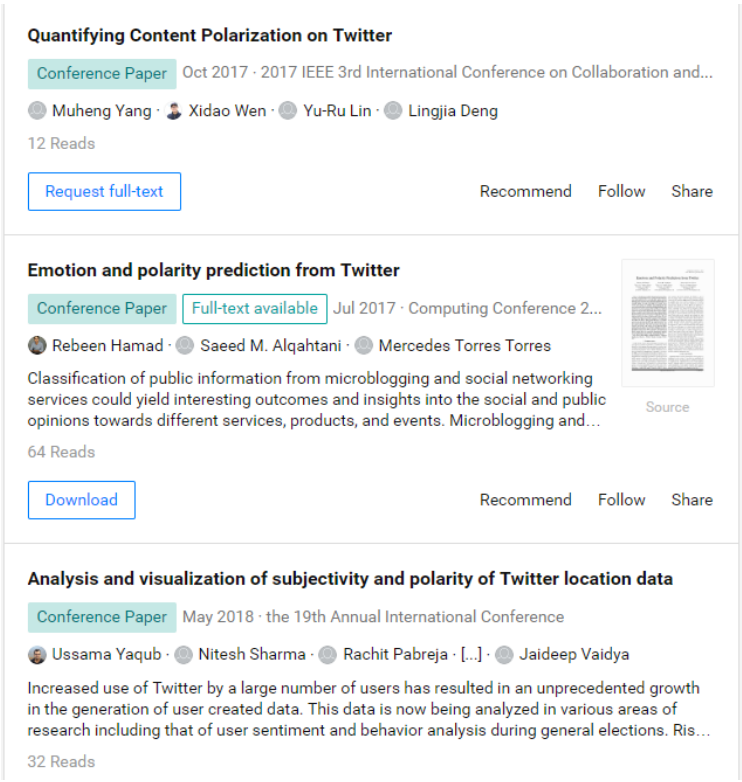
Figura 9 Búsqueda usando ResearchGate



Fuente (AUTOR, 2018)

Los resultados generados indican los artículos que tienen relación con las palabras clave suministradas.

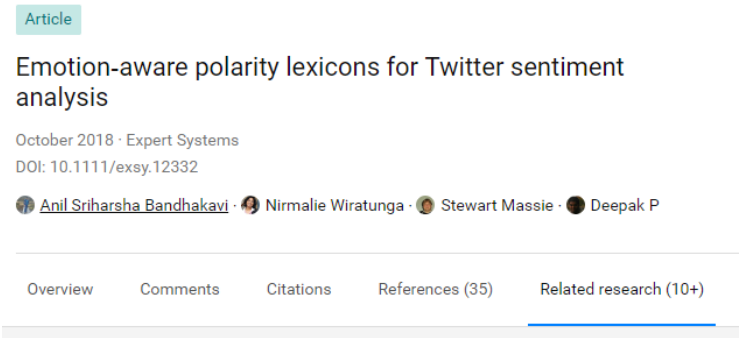
Figura 10 Resultados de búsqueda con ResearchGate



Fuente (AUTOR, 2018)

No solo es posible usar el buscador, también cada artículo tiene una sección con otros artículos relacionados, lo cual puede facilitar la búsqueda de información.

Figura 11 Búsqueda de artículos relacionados



Fuente (AUTOR, 2018)

2.2.1.3 Tabla de selección de algoritmos de datamining

Escoger el mejor algoritmo que se adapte para una tarea en específico no es algo sencillo de realizar, algunas veces se pueden llegar a las mismas conclusiones usando diferentes algoritmos, sin embargo, algunos de ellos pueden generar más de un tipo de resultado. (Microsoft, 2018)

La siguiente tabla permite indicar que clase de técnica podría ser usada según un caso base planteado.

Tabla 1 Tabla de selección de algoritmo de Datamining

Caso base	Técnica sugerida
Predecir un atributo discreto	
Clasificar clientes en una lista de compra como buenos o malos prospectos	Arboles de decisión
Calcular la probabilidad que un servidor falle en los siguientes 6 meses	Naive-Bayes
Categorizar resultados de exámenes y explorar factores relacionados a su diagnostico	Clustering, Redes Neurales
Predecir un atributo continuo	
Pronosticar las ventas del próximo año	Arboles de decisión

Predecir las visitas que tendrá un sitio basado en datos históricos	Regresión Logística
Generar un resultado de riesgo basado en información demográfica	Regresión lineal
Predecir una secuencia	
Analizar un mapa de calor para determinar cuáles son los puntos de interés en una página web	Clustering
Analizar los factores que llevan a que un servidor falle	
Encontrar agrupamientos de ítems comunes	
Sugerir al cliente productos adicionales relacionados a la compra realizada o por realizar	Arboles de decisión
Analizar visitas de un cliente en un evento, para determinar actividades correlacionadas para planear futuras actividades	Arboles de decisión, Redes Neurales
Crear perfiles de riesgo basado en atributos demográficos y comportamientos	Clustering
Analizar usuarios según patrones de compra y navegación	
Identificar servidores que tienen similares características de uso	

Fuente: (Microsoft, 2018)

Dentro de las técnicas mencionadas con anterioridad y siguiendo el caso propuesto indicado como "*Clasificar clientes en una lista de compra como buenos o malos prospectos*", el cual se acopla como concepto base con el objetivo general de este proyecto, se evidencia que la técnica sugerida es ***arboles de decisión***.

2.2.1.4 Procesamiento de Lenguaje Natural

Realizar operaciones analíticas dentro de documentos estructurados puede ser fácilmente ejecutado, sin embargo, no es el mismo caso para documentos no estructurados, como lo son los tweets, es complicado debido a varios problemas como los son ruido digital y datos no especificados. Datamining es otro nombre dado para el análisis de sentimientos (Siddhart, R.Darsini, & Sujithra, 2018, p. 2). En muchas áreas, como negocios, política y acciones públicas, determinar el análisis de sentimiento es importante para entender el sentimiento de la persona o el consumidos de contenido para desarrollar nuevas ideas o continuar con las actividades propuestas.

Algunos políticos de renombre mundial como lo son Barack Obama, Donald Trump y algunos con poca experiencia, usan twitter para desarrollar un vínculo con sus seguidores. También ayuda a aumentar la cantidad de donaciones percibidas para sus campañas. (Petrova, Sen, & Yildrim, 2016, p. 2)

Hay cuatro categorías en las cuales se puede hacer la clasificación de sentimientos usando aprendizaje de maquina:

- Procesamiento de lenguaje natural (PLN): Artificial tiene como propósito ayudar a entender, interpretar y manipular el lenguaje humano, de ese modo ayuda a cerrar la barrera de comunicación entre hombre y máquina.

  Este tipo de procesamiento, tiene en cuenta varias técnicas usadas en datamining para la interpretación del lenguaje humano, partiendo desde métodos estadísticos, procesos de aprendizaje de máquina, enfoques algorítmicos y demás, es decir, condensa algunos de los métodos que componen el ecosistema de minería de datos para generar un modelo de análisis de texto, teniendo en cuenta no solo el significado de cada palabra, sino su contexto y morfología. (SAS, 2018).

- Clasificación estadística: Tienen eficiencia computacional (métodos univariados) y son usados para realizar operaciones donde es necesario tener un resultado del análisis realizado.
- Clustering: Es usada para hacer clasificaciones a alto nivel, es adecuado para extracción de características primarias, pero uno de los inconvenientes que tiene es la extracción de características secundarias.
- Hibrida: Son el resultado de la combinación de métodos univariados y multivariados para obtener resultados acertados.

De acuerdo al enfoque propuesto (Siddhart, R.Darsini, & Sujithra, 2018, p. 3), los casos de clasificación que encajan con la posible resolución del objetivo principal del documento son **Procesamiento de Lenguaje Natural** e **Hibrida**.

2.2.2 Reúso

En la fase de recuperación, las tres técnicas identificadas fueron:

- Arboles de decisión
- Análisis de sentimientos vía PLN
- Hibrida

El PLN reúne varias técnicas para procesar la información de un caso (Nelson, Natural Language Processing (NLP) Techniques for Extracting Information, 2018), por lo tanto, las anteriormente descritas están dentro de este contexto, sin embargo no todos los proyectos son factibles para implementación de PLN. La siguiente imagen describe un diagrama de flujo con algunas decisiones que permiten decidir la factibilidad de aplicación de NLP a una solución propuesta.

2.2.3 Revisión

El tercer objetivo específico de este proyecto (ver DESARROLLAR UNA PRUEBA DE CONCEPTO WEB, QUE TENGA EN CUENTA LOS MÓDULOS DEFINIDOS EN LA ESTRUCTURA DEL MODELO) contempla la implementación de los conceptos identificados en la fase de recuperación y reúso.

2.2.4 Retención

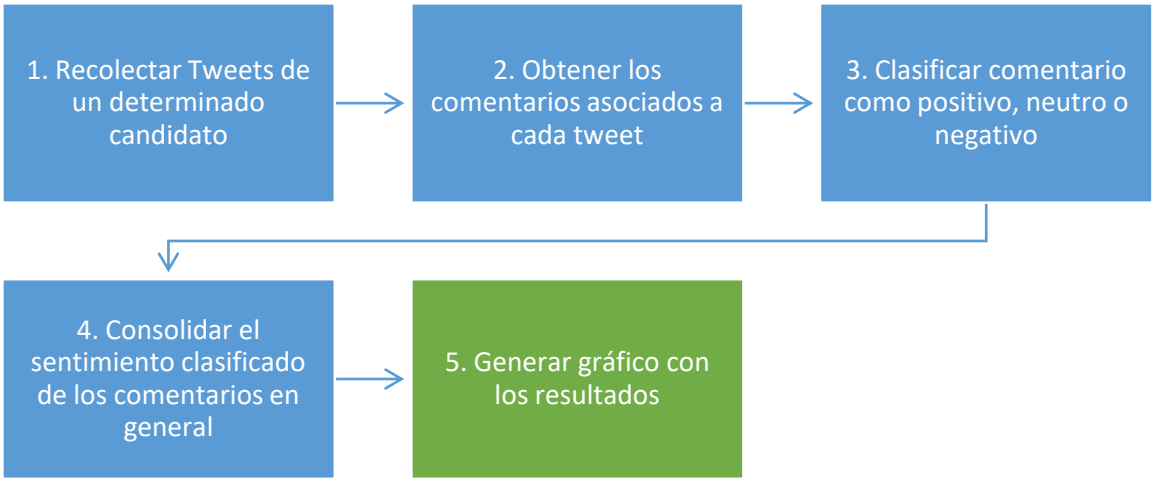
La clasificación en la biblioteca de la Universidad Católica de Colombia de este proyecto, permitirá que futuras investigaciones relacionadas al área de polarización de comentarios en twitter puedan referenciar el trabajo y realizar extensiones o mejoras al mismo o a otros proyectos similares.

3. DISEÑAR LA ESTRUCTURA DEL MODELO RESPECTO AL RESULTADO DE LA FASE DE ANÁLISIS

La metodología usada para plantear el modelo es KDD, el cual define una serie de pasos para plantear un modelo que permita describir los patrones encontrados desde el planteamiento del problema hasta la consolidación de resultados obtenidos.

El esquema del modelo se muestra en la figura 11

Figura 12 Diagrama del proceso del modelo



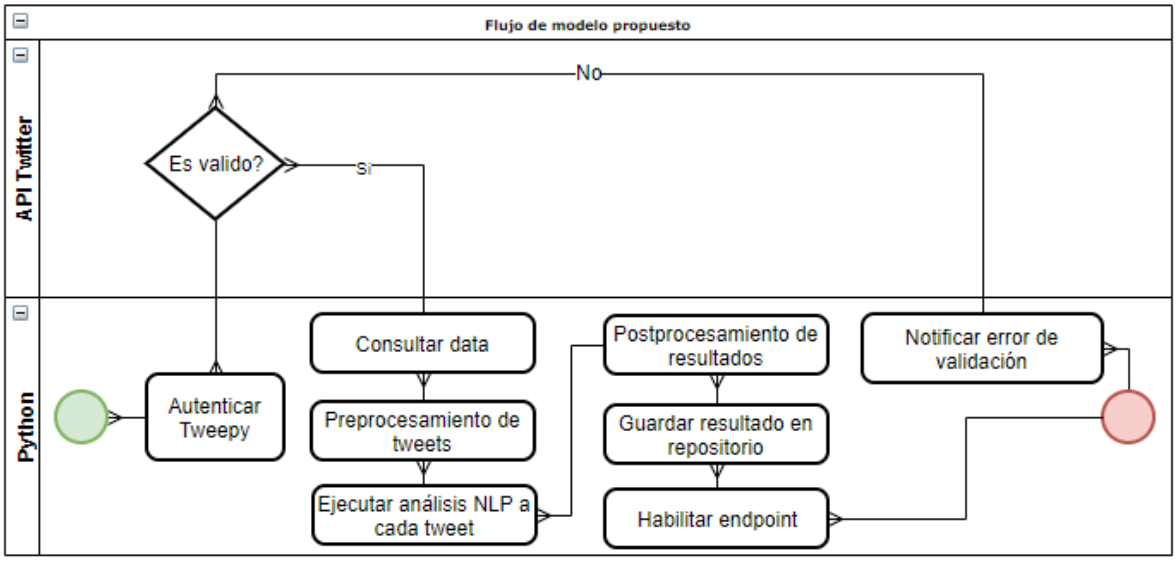
Fuente (AUTOR, 2018)

3.1 Entendimiento del dominio, glosario y conceptos clave

El entendimiento del dominio se refiere a la comprensión que se tiene respecto al problema planteado. En la justificación del proyecto, glosario y en el desarrollo del mismo se expresan los conceptos clave que componen la solución descrita en este documento.

La figura 12 indica el flujo de proceso bajo el cual funciona el modelo propuesto.

Figura 13 Flujo general del modelo propuesto



Fuente (AUTOR, 2018)

3.2 Determinar variables

La tabla 2 muestra el listado de variables usadas dentro del modelo para configurar la ejecución del flujo de proceso del mismo.

Tabla 2 Variables del modelo

Variable	Tipo	Descripción	Ejemplo
Usuario	Cadena(50)	Nombre del usuario en twitter con el cual se extrae la data	@IvanDuque
IncluirRetweets	Booleano	Indicador para retornar tanto las publicaciones propias como los retweets del usuario de twitter.	True/False
RetornarTweetExtendido	Booleano	Indicador para retornar el tweet completo o solo una parte de él.	True/False
ConsumerKey	Cadena(25)	Cadenas que permiten la comunicación hacia la Api de twitter por medio de algún cliente. Son generadas por la plataforma para desarrolladores de twitter	
ConsumerSecret	Cadena(50)		
AccessToken	Cadena(50)		
AccessTokenSecret	Cadena(50)		
Tweet	Cadena(280)	Representación de una publicación generada por un usuario en twitter, contiene su opinión respecto a un tema especifico	“Cuál es la edad aproximada del planeta tierra?”
Comentario	Cadena(280)	Representación de la reacción de un seguidor del usuario ante una publicación	“4.5 billones de años”

Fuente (AUTOR, 2018)

3.3 Recopilación de fuente de datos

Los datos a los cuales se les realiza el minado, son consultados por medio del endpoint GET *statuses/user_timeline*, disponible en el API de twitter. Esta operación retorna los últimos 3200 tweets generados por el usuario en consulta y por medio de una rutina de código, son guardados en un repositorio, el cual será reducido, limpiado y preprocesado antes de su respectivo análisis.

El endpoint está ubicado en *https://api.twitter.com/1.1/statuses/user_timeline.json*. Las tablas 3 y 4 muestran la información de los parámetros de entrada y de salida relevantes para el desarrollo de la solución.

3.3.1 Datos de entrada

Tabla 3 Recopilación - Datos de entrada

Nombre	Tipo	Descripción	Valor por defecto
screen_name	Cadena	Identificador visual del usuario del cual se hará la extracción de tweets.	
count	Entero	Conteo de tweets a retornar, para este caso, siempre será de 200, respectivo al número máximo permitido por consulta al API.	200
tweet_mode	Cadena	Indica el tipo de tweet a retornar.	“Extended”
include_rts	Booleano	Indica si trae los retweets o solo publicaciones hechas por el usuario.	False
id	Numero	Identificador del tweet. Se usa para extraer los comentarios del mismo.	

Fuente (AUTOR, 2018)

3.3.2 Datos de salida

Tabla 4 Recopilación - Datos de salida

Nombre	Tipo	Descripción
full_text	Cadena	Tweet completo
Id	Numero	Identificador del tweet
retweeted_status	Arreglo	Arreglo que contiene los comentarios generados por la comunidad de seguidores a un tweet en particular.

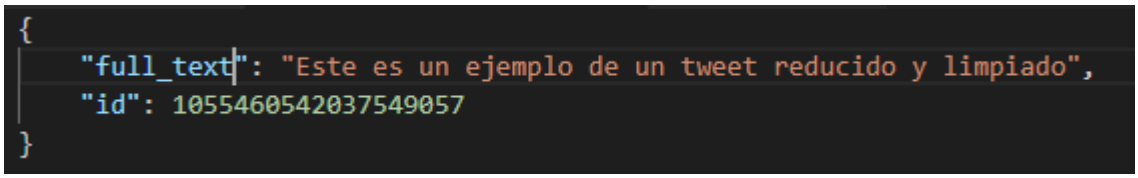
Fuente (AUTOR, 2018)

3.4 Reducción de datos, limpieza y preprocesamiento

La información retornada por el API de twitter contiene bastantes campos con relevancia para proyectos de distintos tipos, tales como geolocalización, marketing digital o publicidad entre otros. Para el desarrollo del objetivo general propuesto, solo es relevante tomar el texto completo del tweet junto a sus comentarios asociados.

Un tweet y comentario limpio son aquellos que contienen solo letras, números y espacios, se eliminan vínculos y caracteres especiales. La Figura 14 muestra un tweet preprocesado con id y texto completo.

Figura 14 Tweet preprocesado y limpio



Fuente (AUTOR, 2018)

3.5 Seleccionar la técnica de datamining

Esta fase hace referencia al primer objetivo específico del documento. La técnica sugerida es PLN – Minería de opinión.

3.6 Datamining

La minería de opinión, permite evaluar la polaridad de un comentario, de ese modo se podría tener un criterio general de la opinión de los seguidores de un usuario respecto a una publicación generada por el mismo.

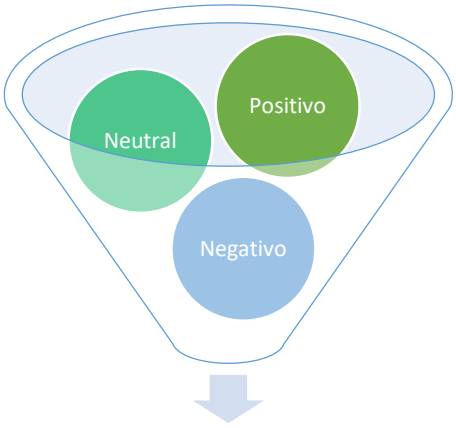
La polaridad está en el rango de [0, 1], lo cual indica el carácter del comentario, los criterios para determinar si es negativo, neutro y positivo están en la tabla X

Tabla 5 Equivalencias de polaridad

Polaridad	Polaridad menor a 0.4	Polaridad entre 0.4 y 0.6	Polaridad mayor a 0.6
Carácter	Negativo	Neutro	Positivo

Fuente (AUTOR, 2018)

Los resultados de la polarización indican el carácter de la publicación.



Caracter de la publicación

Fuente (AUTOR, 2018)

Este proceso, debe realizarse a cada uno de los comentarios del tweet, así una vez se tenga la polaridad calculada de cada uno, se puede hallar el conteo de cuantos elementos negativos, neutros y positivos que determina el carácter general de la publicación.

3.7 Interpretar resultados

Cada tweet de cada usuario estará asociado con la polaridad calculada de cada comentario relacionado, de modo que se pueda realizar el conteo de cada polaridad y sea posible determinar si hay más comentarios positivos, neutrales o negativos en torno a la opinión publicada.

La tabla 6 muestra un ejemplo de un tweet junto a la polaridad calculada de cada comentario. Como resultado es posible determinar que el tweet de seis opiniones, tiene cuatro positivas, una neutral y otra negativa, por lo tanto, la opinión compartida es positiva.

Tabla 6 Ejemplo de tweet, comentarios y caracter

Usuario	Tweet	Comentarios	Carácter
usuario1	Acabo de comprar una guitarra, cual canción me recomiendan aprender?	Me alegra que aprendas a tocar un instrumento, aprende “escalera al cielo”	Positivo
		La que quiera, da lo mismo	Neutral
		Ese instrumento es aburrido y simple, debería aprender otra cosa.	Negativo
		¿Cuál guitarra compraste? Me gustan las ibanez. Comienza con canciones de nirvana	Positivo
		Sería interesante que aprendieras alguna obra de Paco de Lucia. Animo	Positivo
		Trae tu guitarra a mi casa y te enseño algunas canciones, podríamos comenzar nuestro proyecto de la banda.	Positivo

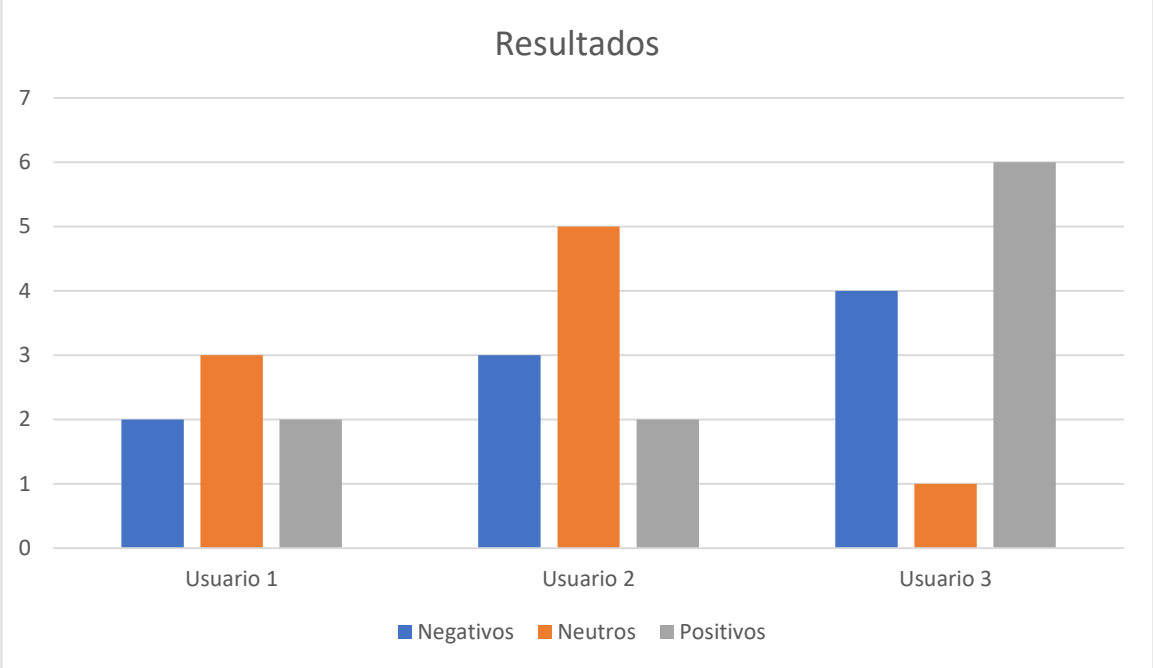
Fuente (AUTOR, 2018)

3.8 Consolidar conocimiento descubierto

En un gráfico, es posible consolidar la polarización de los tweets analizados, de modo que se pueda evidenciar que tan aceptado o rechazado es un candidato basado en las reacciones de sus seguidores a sus publicaciones.

La figura 15 muestra un ejemplo de polarización calculada para tres usuarios, lo cual indica que “Usuario 3” es quien tiene la mayor aceptación debido a su alto número de valoraciones positivas en contraste de las neutras y negativas, mientras que “Usuario 2” tiene una calificación más neutral.

Figura 15 Resultados consolidados



Fuente (AUTOR, 2018)

4. DESARROLLAR UNA PRUEBA DE CONCEPTO WEB, QUE TENGA EN CUENTA LOS MÓDULOS DEFINIDOS EN LA ESTRUCTURA DEL MODELO

4.1 Propósito

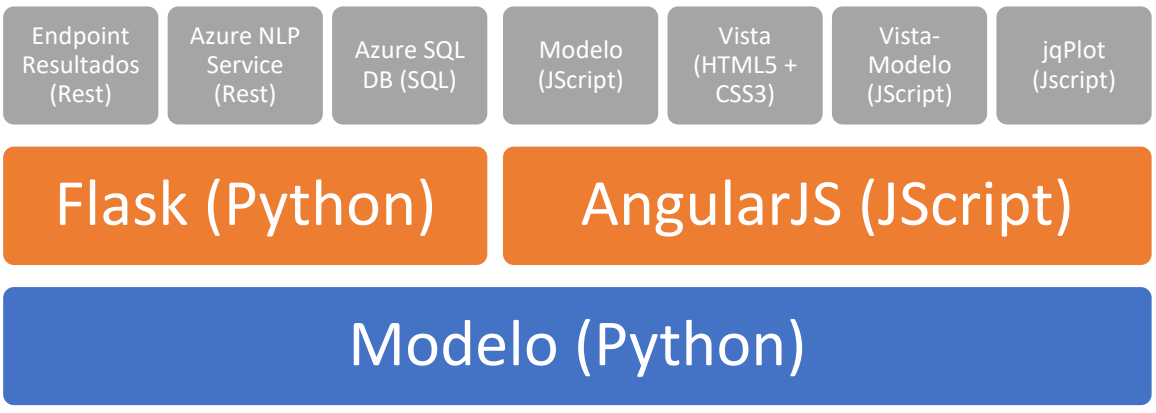
Una de las ventajas que ofrece generar un sitio web, es la reducción de dependencias por parte del cliente al momento de consumir el contenido expuesto, de modo solo necesite usar un navegador para acceder a la página.

La prueba de concepto planteada permitirá a un usuario visualizar los resultados obtenidos por el análisis de los comentarios de forma intuitiva y rápida, al igual los resultados serán generados en formato JSON por medio de un endpoint usando servicios Rest, permitiendo que aplicaciones hechas en otros lenguajes y/o arquitecturas puedan usar la información expuesta para bien sea extenderla o para mostrarla.

4.2 Planteamiento

Para el desarrollo de un sitio web, se necesitan básicamente dos componentes: Un servidor de aplicaciones y el contenido a mostrar, sin embargo, es necesario seguir un patrón de arquitectura para permitir la futura y/o posible extensión del sistema usando un mismo esquema. La figura 16 muestra los componentes propuestos para el levantamiento del sitio web.

Figura 16 Planteamiento técnico de la PDC



Fuente (AUTOR, 2018)

4.2.1 Dependencias

Se decide usar Python 3.7 como lenguaje base del proyecto, ya que, gracias a su amplio catálogo de paquetes, permite realizar sitios web con facilidad. Además, el lenguaje es de fácil uso, sintaxis sencilla y puede correr en distintos sistemas operativos.

Los paquetes descritos en la tabla 7, se refieren a los componentes que la prueba de concepto usa para desplegar los resultados del análisis a los comentarios tanto en el sitio web como en el servicio Rest expuestos.

Tabla 7 Dependencias

Paquete	Descripción
tweepy	Librería que encapsula los llamados a la API de twitter de modo que pueda ser invocado como métodos en Python
pyodbc	Permite la conexión entre un script hacia una base de datos, al igual que la gestión de los resultados generados por la ejecución de consultas
requests	Permite realizar peticiones HTTP sin necesidad de tener un navegador instalado. Es bastante útil para consumir servicios Rest
flask	Es un microframework para hacer fácil y rápidamente aplicaciones web o servicios Rest, Al instalarlo por primera vez, provee un servidor de aplicaciones simple, lo cual para la prueba de concepto es suficiente ya que no se necesitan más componentes
angularJS	Es un framework para crear aplicaciones web dinámicas usando el patrón de arquitectura MVVM (Modelo, Vista, Vista-Modelo). Permite separar claramente las vistas (paginas HTML), los modelos (archivos jscript referentes a controladores) y Vistas-Modelo (Enlaces de datos), lo cual se complementa con el servidor de aplicaciones ofrecido por Flask ya que no se usará algún complemento para renderizado de páginas vía servidor
jqPlot	Componente para generación de gráficas usando colecciones de javascript. Es usado para armar el gráfico de barras que condensa los resultados del análisis de los tweets
bootstrap	Framework CSS para generación de sitios web responsivos. Se usa para llevar orden en la generación y reúso de estilos tanto de la plantilla general del sitio como para sus componentes

Fuente (AUTOR, 2018)

Para calcular la polaridad de un texto y guardarlo en un repositorio, se usan dos servicios en la nube ofrecidos por Azure, TextMining y SQL Database. De ese modo no hay que montar algún servidor adicional para realizar esas tareas.

Tabla 8 Dependencias usando servicios en la nube

Servicio	Descripción
Azure TextMining	Endpoint Rest que le da una puntuación entre 0 y 1 a un texto, indicando aquellos valores cercanos a 0 como negativos, entre 0.4 y 0.6 como neutros y los cercanos a 1 como positivos.
Azure SQL Database	Servicio de hosting de bases de datos. Permite acceder a un repositorio SQL sin necesidad de montar un servidor desde ceros.

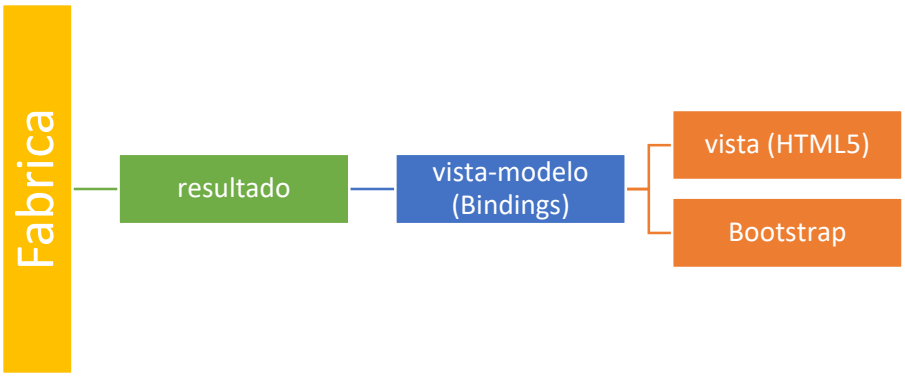
Fuente (AUTOR, 2018)

4.3 Implementación

4.3.1 Capa frontend

Se usa AngularJS para definir una fábrica de gestión de llamados a los servicios Rest generados por flask, estos están inyectados al controlador encargado de mostrar y transformar los resultados. La prueba de concepto solo requiere de tres

Figura 17 Implementación capa frontend



Fuente (AUTOR, 2018)

4.3.2 Capa backend

Se divide el problema en los archivos indicados en la tabla 9.

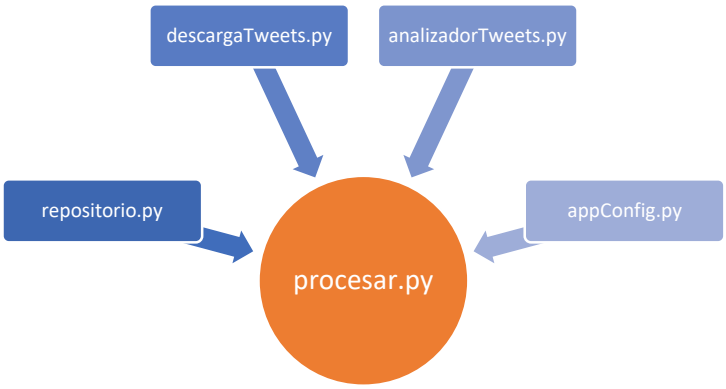
Tabla 9 División de problemas en backend

Archivo	Descripción
app.py	Contiene la configuración de flask y los enrutamientos definidos para exponer información como servicios Rest
appConfig.py	Permite indicar cuales son los usuarios de twitter a los cuales se les realizará el procesamiento de comentarios
analizadorTweets.py	Encapsula el llamado a la API de análisis de texto de Azure, es quien determina la polaridad de un texto
descargaTweets.py	Contiene los métodos usados con tweepy para descarga y limpieza de tweets y comentarios
procesar.py	Script que orquesta los llamados a los métodos expuestos por los otros scripts. Ejecuta el proceso completo desde descarga hasta consolidación de resultados. Debe ser invocado desde consola.
repositorio.py	Gestiona los procesos de guardado, consolidación y retorno de resultados en el repositorio

Fuente (AUTOR, 2018)

La figura 18 describe como convergen los scripts mencionados en la tabla 9 hacia un único punto (procesar.py)

Figura 18 Implementación capa backend



Fuente (AUTOR, 2018)

Los resultados de la polarización de tweets se exponen en un endpoint Rest para que puedan ser consultados bien sea por el modelo de angularJS o por un cliente externo. El objetivo es dejar abierto los resultados para que puedan ser procesados no solo por la prueba de concepto, sino por otros clientes que tengan interés en realizar procesos de análisis de datos.

Figura 19 Retorno de resultados vía endpoint por app.py

Fuente (AUTOR, 2018)

4.3.3 Capa datos

Los resultados de la polaridad de cada comentario son almacenados en una base de datos, de ese modo usando un procedimiento almacenado, es posible realizar un ETL que consolide esos resultados a una tabla que será el suministro de información al endpoint Rest expuesto.

El modelo expuesto en la figura X, indica las 2 tablas que se usan para guardar y consolidar los resultados. La tabla *TweetsPolarizados* almacena cada comentario asociado a un tweet con su respectivo puntaje calculado. La tabla *ConsolidadoPolarizacion* guarda los resultados calculados por el procedimiento *PA_ConsolidarResultados*, el cual por medio de expresiones comunes de tabla realiza los conteos de cuantos tweets negativos, neutros y positivos hay.

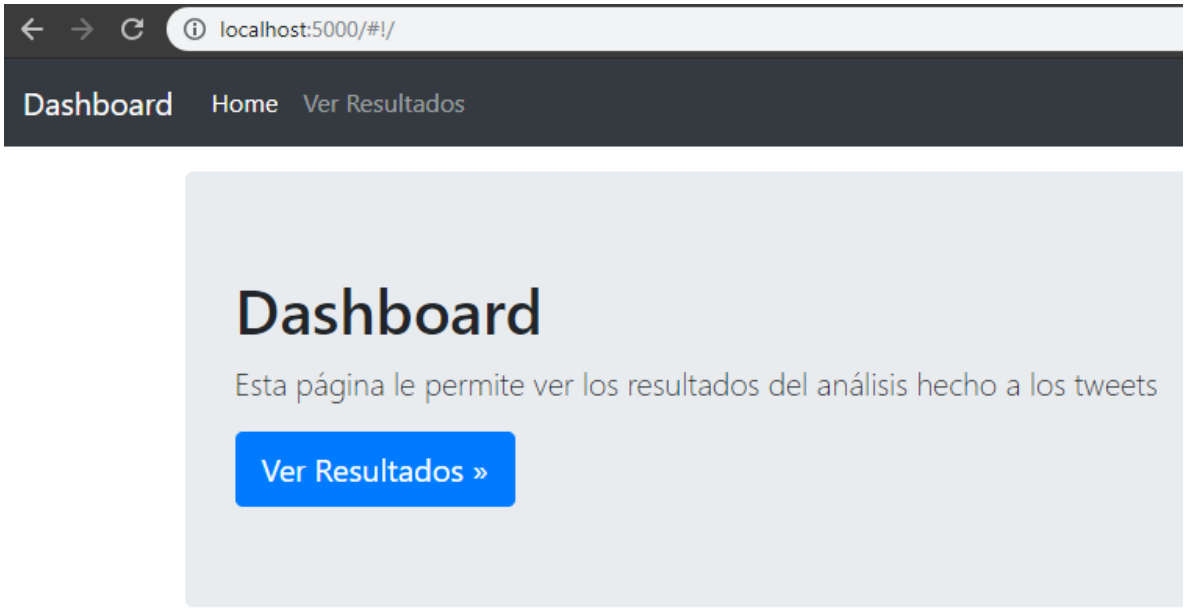
Figura 20 Diagrama entidad-relación

Fuente (AUTOR, 2018)

4.4 Resultados

Ejecutando el script app.py, se levanta un servidor de aplicaciones ligero corriendo por puerto 5000 en el host local, lo cual usando el enrutamiento de angularJS procesa la plantilla general y muestra un mensaje descriptivo de lo que puede consultar el cliente.

Figura 21 Aplicación ejecutada en navegador.

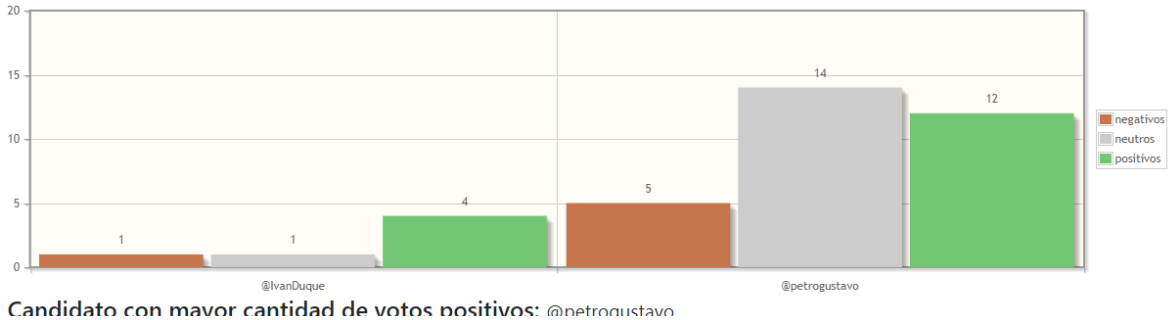


Fuente (AUTOR, 2018)

El acceso a resultados, muestra una gráfica con 3 barras, las cuales significan

- Rojo: Tweets negativos.
- Gris: Tweets neutros.
- Verde: Tweets positivos.

Figura 22 Resultados de la PDC



Fuente (AUTOR, 2018)

5. EJECUTAR UN CONJUNTO DE PRUEBAS FUNCIONALES QUE PERMITAN EVALUAR LA CALIDAD DE LOS RESULTADOS CALCULADOS POR LA SOLUCIÓN

Las pruebas funcionales permiten validar el resultado de las salidas de un proceso bajo unos parámetros de entrada específicos, en busca que lo implementado cumpla con la función especificada. Estas son comúnmente identificadas como pruebas de caja negra.

5.1 Requisitos

Para la ejecución de los planes de pruebas, es necesario contar con los recursos descritos en la tabla 10.

Tabla 10 Requisitos para pruebas funcionales

Recurso	Descripción	Parametrización requerida
Testfile para Azure DB (testDB.py)	Archivo que contiene parametrización fija que permite verificar que la base de datos está operativa.	* Nombre del servidor * Usuario * Clave * Nombre de la BD
Testfile para Azure TextAnalytics (testAnalizador.py)	Archivo con parametrización fija que ejecuta un análisis a varios textos en busca de calcular la polaridad de los mismos	* Endpoint * Key OCP
Testfile para Tweepy (testTweepy.py)	Archivo con parametrización fija, el cual permite verificar que la conexión a la API de twitter es correcta, al igual que la descarga de tweets y comentarios relacionados	* Llave consumidor * Clave consumidor * Llave acceso * Clave acceso

Fuente (AUTOR, 2018)

5.2 Plan de pruebas

Una vez cubiertos los requisitos, se puede usar una tabla que describe la funcionalidad evaluada, código de la prueba ejecutada, variables de entrada, resultado y observaciones.

Tabla 11 Ejemplo formato tabla plan de pruebas

Funcionalidad evaluada	Código de la prueba	Variables de Entrada	Resultado	Observaciones	Aceptada?
Recurso 1	X001	Param1: Val1 Param2: Val2 Param3: Val3	Resultado obtenido	Observación (si aplica)	Si/No

Fuente (AUTOR, 2018)

Sin embargo, para la definición de las pruebas, no es necesario tener valores asociados en las columnas “Resultado”, “Observaciones” ni “Aceptada”, ya que son usadas al recibir los resultados de ejecución de la misma.

5.2.1 Pruebas de polarización

Estas pruebas permiten evaluar los resultados generados usando el recurso *testAnalizador.py* respecto a los resultados del API de análisis de texto de Azure. La respuesta del API se encuentra en el anexo A (ver Anexo A),

5.2.1.1 Validar Positivismo

Las pruebas expresadas en la tabla 12 determinan si los resultados generados por el Api de análisis de texto de Azure retornan puntajes dentro del rango considerado como positivo con los textos provistos.

Tabla 12 Pruebas para puntajes positivos

Funcionalidad evaluada	Código de la prueba	Variables de Entrada
Validar puntaje de positivismo retornado por el API de análisis de texto de Azure	P010	Texto: “*Estoy muy feliz*”
	P011	Texto: “*El nuevo aeropuerto esta increíble, es muy grande y tiene muchos locales, me siento cómodo*”
	P012	Texto: “si quieres, puedes hacerlo, estoy seguro que puedes lograrlo, con esfuerzo lo conseguirás!”

Fuente (AUTOR, 2018)

5.2.1.2 Validar Neutralidad

Las pruebas expresadas en la tabla 13 determinan si los resultados generados por el Api de análisis de texto de Azure retornan puntajes dentro del rango considerado como neutro con los textos provistos.

Tabla 13 Pruebas para puntajes neutros

Funcionalidad evaluada	Código de la prueba	Variables de Entrada
Validar puntaje de neutralidad retornado por el API de análisis de texto de Azure	P020	Texto: "*me da igual*"
	P021	Texto: "*de acuerdo, sin embargo no lo hagas*"
	P022	Texto: "*meh, si, como quieras*"

Fuente (AUTOR, 2018)

5.2.1.3 Validar Negatividad

Las pruebas expresadas en la tabla 14 determinan si los resultados generados por el Api de análisis de texto de Azure retornan puntajes dentro del rango considerado como negativo con los textos provistos.

Tabla 14 Pruebas para puntajes negativos

Funcionalidad evaluada	Código de la prueba	Variables de Entrada
Validar puntaje de negatividad retornado por el API de análisis de texto de Azure	P030	Texto: "Estoy muy triste"
	P031	Texto: "*tengo muchos problemas, no estoy seguro de lo que debo hacer*"
	P032	Texto: "*hoy me levante con ganas de llorar*"

Fuente (AUTOR, 2018)

5.3 Ejecución

Basado en los casos de prueba propuestos, se parametrizan los diccionarios que procesará el API de análisis de texto (ver Anexo A). Allí se indica el valor exacto del puntaje de polaridad asignado a cada prueba, para efectos de visualización en las tablas 15, 16 y 17 se usan solo dos dígitos decimales.

5.3.1.1 Validar Positivismo

Tabla 15 Resultados de la prueba para comentarios positivos

Código de la prueba	Variables de Entrada	Resultado	Observaciones	Aceptada?
P010	Texto: "*Estoy muy feliz*"	0.81		Si
P011	Texto: "*El nuevo aeropuerto esta increíble, es muy grande y tiene muchos locales, me siento cómodo*"	0.77		Si
P012	Texto: "si quieres, puedes hacerlo, estoy seguro que puedes lograrlo, con esfuerzo lo conseguirás!"	0.75		Si

Fuente (AUTOR, 2018)

5.3.1.2 Validar Neutralidad

Tabla 16 Resultados de la prueba para comentarios neutros

Código de la prueba	Variables de Entrada	Resultado	Observaciones	Aceptada?
P020	Texto: "*me da igual*"	0.47		Si
P021	Texto: "*de acuerdo, sin embargo no lo hagas*"	0.40		Si
P022	Texto: "*meh, si, como quieras*"	0.63	Esta por fuera del rango definido como neutro	No

Fuente (AUTOR, 2018)

5.3.1.3 Validar Negatividad

Tabla 17 Resultados de la prueba para comentarios negativos

Código de la prueba	Variables de Entrada	Resultado	Observaciones	Aceptada?
P030	Texto: "Estoy muy triste"	0.0		Si
P031	Texto: "*tengo muchos problemas, no estoy seguro de lo que debo hacer*"	0.38		Si
P032	Texto: "*hoy me levante con ganas de llorar*"	0.05		Si

Fuente (AUTOR, 2018)

6. CONCLUSIONES

Las redes sociales no son solo herramientas donde la gente interactúa e intercambia su opinión respecto a un tema, estas evolucionaron para ser plataformas donde se puede influir la percepción que un usuario tiene sobre algún tema en particular. También se presenta como una herramienta decisiva en la toma de decisiones, ya que, compartiendo una opinión o una idea, se podría inferir lo que piensa el público objetivo.

El modelo propuesto permitió consolidar la polarización de varias publicaciones realizadas por los candidatos a la presidencia de Colombia para el periodo 2018-2022, de ese modo fue posible determinar qué tan aceptado o rechazado podría ser de acuerdo al sentimiento expresado en los comentarios por sus seguidores, sin embargo, uno de los temas a mejorar es la detección del sarcasmo que usan algunos usuarios al comentar un tweet, además de la inclusión de más portales (redes sociales, noticias, datos abiertos, datos enlazados).

Estos elementos generaron algunos resultados clasificados como falsos positivos, donde el sentimiento real del sarcasmo es negativo pero la frase analizada quedaba con puntaje positivo, lo cual alteró los datos consolidados consumidos posteriormente por el API de resultados.

La exposición de los resultados como un endpoint Rest, permitió que clientes para prueba de peticiones HTTP hechos en otras tecnologías y corriendo sobre sistemas operativos diferentes, pudieran acceder a los datos consolidados de la polarización. Este concepto fue útil para extender la solución.

7. TRABAJOS FUTUROS

- Incluir más redes sociales, tales como Facebook o YouTube, ya que los candidatos tienen una fuerte presencia en esos canales y muchas personas opinan y comparten estados por esos portales.
- Incluir análisis audiovisual. Algunas ocasiones los candidatos publican imágenes o videos que son generadores de opiniones diversas.
- Tener en cuenta los emoticonos al momento de realizar la minería de opinión, estos pueden maximizar o minimizar el sentimiento con el cual fue escrito un comentario.
- Procesar el sarcasmo indicado por los comentarios de un seguidor hacia una publicación.
- Usar otros enfoques de minería de opinión, usando un proceso local con librerías tales como NLTK, spaCy, Gensim u otras en vez de consumos hacia una API de Azure, la cual tiene límites de peticiones por hora en su versión gratuita.
- Usar la API Premium de Twitter para no tener restricciones de tiempo ni cuotas al momento de realizar búsquedas.
- Tomar en cuenta las publicaciones de los influenciadores políticos.

BIBLIOGRAFÍA

aaai.org. (1995, 08 20). *The first international conference on knowledge discover and data mining*. (AAAI) Retrieved from https://aaai.org/Conferences/KDD/kdd95.php

Akhilomen, J. (2013). *Data Mining Application for Cyber Credit-Card Fraud Detection System*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-39736-3_17

Alang, N. (2016, Nov 15). *Trump Is America's First Twitter President. Be Afraid.* Retrieved from https://newrepublic.com/article/138753/trump-americas-first-twitter-president-afraid

Aldrich, J., & Garrod, C. (2013). *Principles of Software Construction: Conceptual Modeling in Design*. (ISR Institute for software research) Retrieved from Carnegie Mellon University: https://www.cs.cmu.edu/~aldrich/214/slides/conceptual-modeling.pdf

Amazon. (2018). *Data Warehouse Concepts*. Retrieved 10 2018, from Amazon: https://aws.amazon.com/data-warehouse/

Andrew Lampitt. (2013, Febrero 14). *The real story of how big data analytics helped Obama win*. (infoworld) Retrieved from https://www.infoworld.com/article/2613587/big-data/the-real-story-of-how-big-data-analytics-helped-obama-win.html

Ardila, C. (2018). *Modelado del Dominio(Modelo conceptual)*. (Universidad del Cauca) Retrieved from Universidad del Cauca: http://artemisa.unicauca.edu.co/~cardila/IS_04_03__MODELADO_DOMINIO_NV.pdf

AUTOR, E. (2018). *Tecnicas de Datamining.* Recuperado el 01 de 06 de 2018

Będkowski, K. (2011, 01 01). *Trees crowns segmentation on the basis of a digital surface model obtained from the interpolation of airborne laser scanning data*. Retrieved from ResearchGate: https://www.researchgate.net/publication/306374119_Trees_crowns_segmentation_on_the_basis_of_a_digital_surface_model_obtained_from_the_interpolation_of_airborne_laser_scanning_data

Berzal, F. (2017, 01 01). *Clústering Jerárquico*. (Universidad de Granada) Retrieved from http://elvex.ugr.es/idbis/dm/slides/42%20Clustering%20-%20Hierarchical.pdf

Bhattacharyya, S. (2011, 02). *Data mining for credit card fraud: A comparative study*. (ScienceDirect) Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923610001326

Bouza, C., & Santiago, A. (2014, 10 21). *LA MINERÍA DE DATOS: ARBOLES DE DECISIÓN Y SU APLICACIÓN EN ESTUDIOS MÉDICOS*. Retrieved from ResearchGate: https://www.researchgate.net/publication/268516570_LA_MINERIA_DE_DATOS_ARBOLES_DE_DECISION_Y_SU_APLICACION_EN_ESTUDIOS_MEDICOS

Clifton, C. (2018, 08 01). *Data Mining, Computer Science*. (Enciclopedia Britannica) Retrieved from https://www.britannica.com/technology/data-mining

CrayonData. (2015, 04 1). *What is Clustering in Data Mining?* Retrieved from http://bigdata-madesimple.com/what-is-clustering-in-data-mining/

Dane. (23 de 11 de 2018). *Censo nacional de poblacion y vivienda: ¿Cuantos Somos?* Recuperado el 23 de 11 de 2018, de DANE: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivienda-2018/cuantos-somos

Earls, J. (2010, Febrero 26). *SQL: Query Language for Relational Databases*. (Media College) Retrieved from https://www.mediacollege.com/computer/database/sql.html

Faggella, D. (2017, Sep 2). *What is Machine Learning?* (Techemergence) Retrieved from https://www.techemergence.com/what-is-machine-learning/

Functional testing. (n.d.). (Software Testing Fundamentals) Retrieved from http://softwaretestingfundamentals.com/functional-testing/

Gallagher, S. (2012, 11 14). *Built to win: Deep inside Obama's campaign tech*. (Ars Technica) Retrieved from https://arstechnica.com/information-technology/2012/11/built-to-win-deep-inside-obamas-campaign-tech/

Gaur, P. (2012). *Neural Networks in Data Mining*. (International Journal of Electronics and Computer Science Engineering) Retrieved from International Journal of Electronics and Computer Science Engineering: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.640.136&rep=rep1&type=pdf

Gibert, K., Sanchez-Marre, M., & Codina, V. (2014, 05 27). *Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation*. Retrieved from ReserachGate: https://www.researchgate.net/publication/228609947_Choosing_the_Right_Data_Mining_Technique_Classification_of_Methods_and_Intelligent_Recommendation

Gil, P. (2018, Feb 05). *What Is Twitter & How Does It Work?* (LifeWire) Retrieved from https://www.lifewire.com/what-exactly-is-twitter-2483331

Gupta, P. (2017, 05 17). *Decision Trees in Machine Learning*. (Towards Data Science) Retrieved from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Hernández-Orallo, J. (2015). *Visualización*. (Universitat Politècnica de València) Retrieved 06 05, 2018, from Universitat Politècnica de València: http://users.dsic.upv.es/~jorallo/docent/doctorat/t2b.pdf

Hui Li. (2017, Abril 12). *Which machine learning algorithm should I use?* Retrieved from SAS: https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

Hutto, C. (2017). *VaderSentiment*. Retrieved from https://github.com/cjhutto/vaderSentiment#about-the-scoring

Jacobs, B. (2012, Nov 11). *Orca Failed; but So Did Obama's 2008 Version of the Same*. (The Atlantic) Retrieved from

https://www.theatlantic.com/politics/archive/2012/11/orca-failed-but-so-did-obamas-2008-version-of-the-same/265077/

Jose Antonio Vargas. (2007, Febrero 17). *Young Voters Find Voice on Facebook*. Retrieved from http://www.washingtonpost.com: http://www.washingtonpost.com/wp-dyn/content/article/2007/02/16/AR2007021602084.html?nav=emailpa

Kissane, D. (2016, Mayo 26). *It's Time To Explain What Facebook Actually Is*. (doz.com) Retrieved from http://www.doz.com/social-media/facebook-explained

Knight, M. (2017, 09 11). *What is Metadata?* (Dataversity) Retrieved from Dataversity: http://www.dataversity.net/what-is-metadata/

Lucidchart. (10 de 2018). *Qué es un diagrama de árbol de decisión*. Obtenido de Lucidchart: https://www.lucidchart.com/pages/es/qu%C3%A9-es-un-diagrama-de-%C3%A1rbol-de-decisi%C3%B3n

Microsoft. (2017, 02 13). *C# Reference*. Retrieved from MSDN Library: https://docs.microsoft.com/es-es/dotnet/csharp/language-reference/index

Microsoft. (2018, 04 30). *Data Mining Algorithms (Analysis Services - Data Mining)*. Retrieved from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017

Nelson, P. (2018, 10). *Deciding If A Natural Language Processing (NLP) Project Is Feasible*. Retrieved from Search Technologies: https://www.searchtechnologies.com/nlp-project-feasibility-flowchart

Nelson, P. (2018, 10). *Natural Language Processing (NLP) Techniques for Extracting Information*. Retrieved from Search Technologies: https://www.searchtechnologies.com/blog/natural-language-processing-techniques

Nickerson, D. W., & Rogers, T. (2013, Noviembre). *Political Campaigns and Big Data.* Retrieved from https://scholar.harvard.edu/files/todd_rogers/files/political_campaigns_and_big_data_0.pdf

Nico, G. (2014, 01 31). *Data Mining - Algorithms*. Retrieved from https://gerardnico.com/_detail/data_mining/data_mining_algorithm.jpg?id=data_mining%3Aalgorithm

Oracle. (2018, Marzo). *What Is Data Mining*. (Oracle) Retrieved from Oracle: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDFGCIJ

Ortiz, Y., Bañuelos, P., & Rodas-Osollo, J. (04 de 2016). Obtenido de Universidad Autonoma de Ciudad Juarez: http://erevistas.uacj.mx/ojs/index.php/culcyt/article/download/1085/951

Petrova, M., Sen, A., & Yildrim, P. (2016, 09 29). *Social Media and Political Donations: New Technology and Incumbency Advantage in the United States*. Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2836323

Ray, S. (2015, 08 14). *7 Types of Regression Techniques you should know!* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

ResolveStudio. (2018). *¿Quién será el Presidente de Colombia en 2018?* Retrieved from http://resolvestudio.co/elecciones-presidenciales-colombia/

Romero-Campero, F. J. (2013). *Búsqueda de patrones: técnicas de clustering*. (Universidad de Sevilla) Retrieved 10 2018, from https://www.cs.us.es/~fran/curso_unia/clustering.html

Rouse, M. (2011, Enero). *MPP (massively parallel processing)*. (techtarget) Retrieved from http://whatis.techtarget.com/definition/MPP-massively-parallel-processing

Rouse, M. (2018, 08). *Facebook definition*. (techtarget) Retrieved Marzo 25, 2018, from WhatIs.com: http://whatis.techtarget.com/definition/Facebook

Sanon, M. (2017, 04 28). *4 Reasons Why Data Analytics is Important*. (Digital Vidya) Retrieved from https://www.digitalvidya.com/blog/reasons-data-analytics-important/

SAS. (2018, 09 30). *Data Mining, What it is and why it matters*. (SAS) Retrieved from https://www.sas.com/en_us/insights/analytics/data-mining.html

SAS. (30 de 09 de 2018). *Natural Language Processing What it is and why matters*. Obtenido de SAS: https://www.sas.com/en_id/insights/analytics/what-is-natural-language-processing-nlp.html

Sasha Issenberg. (2012, Diciembre 19). *How Obama's Team Used Big Data to Rally Voters*. (technologyreview) Retrieved from https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/

Sean Gallagher. (2012, 9 11). *Inside Team Romney's whale of an IT meltdown*. Retrieved from ArsTechnica: https://arstechnica.com/information-technology/2012/11/inside-team-romneys-whale-of-an-it-meltdown/

Semana. (02 de 02 de 2018). *¿'Likes' o votos? Esa es la cuestión en unas elecciones*. Obtenido de Semana: https://www.semana.com/nacion/articulo/campanas-en-redes-sociales/555643

Shethna, J. (2017, Nov 7). *7 Important Data Mining Techniques for Best results*. Retrieved from https://www.educba.com/7-data-mining-techniques-for-best-results/

Siddhart, S., R.Darsini, & Sujithra, M. (2018, 05). *Sentiment Analysis on Twitter Data Using Machine*. Retrieved from ResearchGate: https://www.researchgate.net/publication/325359363

Statcounter. (2018, 11 23). *Social media stats Colombia*. Retrieved from Statcounter GlobalStats: http://gs.statcounter.com/social-media-stats/all/colombia/#monthly-201711-201810

Techopedia. (2015). *Big Data*. (Techopedia Inc) Retrieved Marzo 25, 2018, from Techopedia: https://www.techopedia.com/definition/27745/big-data

The R Foundation. (Marzo de 2018). *What is R?* Obtenido de https://www.r-project.org/about.html

What is extreme programming. (15 de Abril de 2018). (Agile Alliance) Obtenido de https://www.agilealliance.org/glossary/xp/

Why Stata? (2018, Marzo). (StataCorp LLC) Retrieved from https://www.stata.com/why-use-stata/

Zwass, V. (2018, 08 01). *Expert system*. (Enciclopedia Britannica) Retrieved from https://www.britannica.com/technology/expert-system

ANEXOS

Anexo A Json de respuesta API Azure


```
{
    "documents": [
        {"id": "P010", "score": 0.7734057307243347},
        {"id": "P011", "score": 0.7548137307167053},
        {"id": "P012", "score": 0.8139266967773438},
        {"id": "P020", "score": 0.47043874859809875},
        {"id": "P021", "score": 0.4042259454727173},
        {"id": "P022", "score": 0.6312606334686279},
        {"id": "P030", "score": 0.0},
        {"id": "P031", "score": 0.3885572552680969},
        {"id": "P032", "score": 0.055093467235565186}
    ],
    "errors": []
}
```


Anexo B Script generación tabla TweetsPolarizados

```
CREATE TABLE TweetsPolarizados
(
    Id INT IDENTITY(1,1) PRIMARY KEY,
    FechaEjecucion DATETIME NOT NULL DEFAULT GETDATE(),
    Usuario VARCHAR(50) NOT NULL,
    IdTweet VARCHAR(50) NOT NULL,
    Tweet VARCHAR(300) NOT NULL,
    IdComentario VARCHAR(50) NOT NULL,
    Comentario VARCHAR(300) NOT NULL,
    Polaridad NUMERIC(18,2) NOT NULL
)
```

Anexo C Script creación tabla ConsolidadoPolarizacion

```
CREATE TABLE ConsolidadoPolarizacion
(
    Id INT IDENTITY(1,1) PRIMARY KEY,
    Usuario VARCHAR(50) NOT NULL,
    TweetsNegativos INT,
    TweetsNeutros INT,
    TweetsPositivos INT
)
```

Anexo D Script creación SP PA_ConsultarResultados

```
CREATE PROCEDURE PA_ConsultarResultados
AS
BEGIN
    SELECT usuario, tweetsNegativos, tweetsNeutros, tweetsPositivos FROM
ConsolidadoPolarizacion
END
```

Anexo E Script para consolidación de resultados

```
CREATE PROCEDURE PA_ConsolidarResultados
AS
BEGIN
    DELETE ConsolidadoPolarizacion;

    WITH Negativos AS
    (
    SELECT Usuario, Conteo = COUNT(Polaridad), Tipo = 'Negativo' FROM TweetsPolarizados
    WHERE Polaridad < 0.4
    GROUP BY Usuario
    ),
    Neutros AS (
    SELECT Usuario, Conteo = COUNT(Polaridad), Tipo = 'Neutro' FROM TweetsPolarizados
    WHERE Polaridad BETWEEN 0.4 AND 0.6
    GROUP BY Usuario
    ),
    Positivos AS (
    SELECT Usuario, Conteo = COUNT(Polaridad), Tipo = 'Positivo' FROM TweetsPolarizados
    WHERE Polaridad > 0.6
    GROUP BY Usuario
    )
    SELECT
        P.Usuario,
        ConteoPositivos = ISNULL(P.Conteo, 0),
        ConteoNeutros = ISNULL(N.Conteo, 0),
        ConteoNegativos = ISNULL(NG.Conteo, 0)
    INTO #temp
    FROM positivos P
    LEFT JOIN neutros N ON P.Usuario = N.Usuario
    LEFT JOIN negativos NG ON P.Usuario = NG.Usuario

    INSERT INTO ConsolidadoPolarizacion (usuario, tweetsNegativos, tweetsNeutros, tweetsPositivos)
    SELECT usuario, ConteoPositivos, ConteoNeutros, ConteoNegativos FROM #temp
END
```

Anexo F Json formato Python para ejecución de pruebas de polaridad

```
{'documents': [
    {'id': '1', 'language': 'es', 'text': 'estoy muy triste'},
    {'id': '2', 'language': 'es', 'text': 'tengo muchos problemas, no estoy seguro de lo que debo
hacer'},
    {'id': '3', 'language': 'es', 'text': 'hoy me levante con ganas de llorar'},

    {'id': '4', 'language': 'es', 'text': 'me da igual'},
    {'id': '5', 'language': 'es', 'text': 'de acuerdo, sin embargo no lo hagas'},
    {'id': '6', 'language': 'es', 'text': 'meh, si, como quieras'},

    {'id': '7', 'language': 'es', 'text': 'El nuevo aeropuerto esta increíble, es muy grande y tiene
muchos locales, me siento cómodo'},
    {'id': '8', 'language': 'es', 'text': 'si quieres, puedes hacerlo, estoy seguro que puedes
lograrlo, con esfuerzo lo conseguirás!'},
    {'id': '9', 'language': 'es', 'text': 'estoy muy feliz'}
]}
```